Writing across the protein universe

ARRON NATHAN BALE

Writhing across the protein universe

Arron Bale

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematics
Durham University
United Kingdom
September 2024

Abstract

The function of a protein is primarily determined by its specific 3D structure, which is itself informed by the sequence of amino acids that make up the protein. Thus, being able to predict the final 3D shape of a protein from its sequence is of vital importance to researchers. This thesis discusses methods for studying protein structure on various scales, motivated in large part by the development of Carbonara; a software for rapid refinement of protein structure based on experimental solution scattering data. This software fills the gap where machine learning methods such as AlphaFold are unable to produce predictions which accurately capture a proteins structure and dynamics in near native conditions in solution.

The local geometry of a protein is well understood to be tightly constrained by its chemistry, we provide constraints on the super secondary and tertiary scale. To achieve this, we present a novel method for smoothing the protein's backbone curve which produces a minimal representation of the underlying entanglement of its secondary structure elements. By studying the distribution of writhe for these smoothed backbone curves we find clear limits on their entanglement. We show that a large scale helical geometry is responsible for proteins which have maximal entanglement relative to this bound. We show that helical geometries are also dominant as a super secondary motif within proteins, linked to their structural and thermal stability.

We show that there is a clear lower bound on the expected amount of absolute entanglement of the backbone as a function of its secondary structure. This insight was key to the development of Carbonara, with this lower bound acting as a penalty to produce biologically plausible predictions. This is a vital step in Carbonara's pipeline, allowing the coarse grained model to be safely passed into all-atomistic molecular dynamics simulations. We present the framework for a complementary model to Carbonara which uses gradient descent to optimise the backbone curve model.

Declaration

The work in this thesis is based on research carried out at the Department of Mathematics, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

Firstly, thank you to my supervisor Chris Prior for your infectious enthusiasm for our work, your willingness to let me follow my curiosity, and your seemingly laws of physics breaking ability to find time for me despite countless other priorities. I can't believe there's enough coffee in the world to fuel everything you manage to fit into a day.

Secondly, I thank Rob Rambo, for your supervision at Diamond and for the endlessly interesting discussions about proteins, football, self-driving cars, and all sorts in between. My thanks as well to the rest of the team on Beamline 21, I hope you enjoy using our code as much as I enjoyed my visits there.

I must also thank the MoSMed CDT, for having faith in me and this project despite it at times feeling quite far from the rest of your research. Without your support, and the funding from EPSRC and Diamond, none of this work would ever be possible.

To my B, you've been with me throughout my academic growth. In different cities, in different countries, your love has always been the same. It's been a long and difficult journey to get here, but you make everything easier. Thank you.

Finally, I thank my family for your unconditional love and support for me through everything I do. Liam, you're everything an older brother should be, you set the impossibly high, and do everything you can to help me feel like I can reach it. Mum, you taught me what it means to work hard for myself and for others, there aren't enough lifetimes left for me to repay you. Dad, you taught me how to not take any of this too seriously, helping me find the freedom and balance to thrive. After seeing an early iteration of a figure which would prove to be vital to one of my major contributions in this thesis, you joked "is that it?". Well, a few years, a hundred or so pages later, yeah, I guess that's it.

Contents

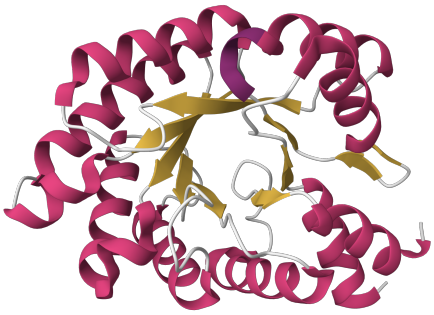
List of Figures

CHAPTER 1

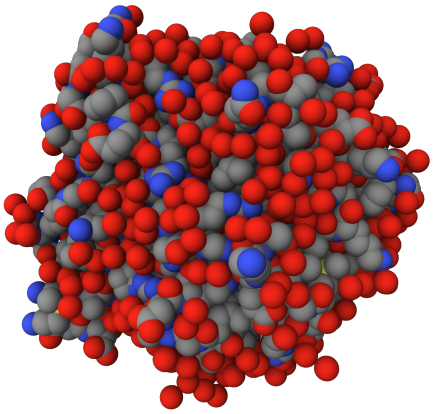
Introduction

1.1 Protein structure

Proteins are the fundamental building blocks of life, playing a role in virtually all biological processes. Some of the first studies on proteins date back to the 18th century, when Antoine Fourcroy (1755-1809) published his work on the separation of gluten from wheat. This work predates the term "protein" itself, which is believed to have been coined by Jacob Berzelius (1779–1848) and Gerrit Mulder (1802-1880) and means "standing in front". By the 1900s, researchers had reached a consensus that proteins are built from amino acids, with Leucine the first to be isolated in 1819 and Threonine the 20th and final of those found in the human body to be isolated in 1936. The first folded 3D structure was published in 1958 by John C. Kendrew et al. For the complete history of protein research in this early period, we recommend [1]. Since the publication of this first structure, the field has advanced massively, with researchers curating the Protein Data Bank (PDB), containing more than 225,000 experimentally determined structures at the time of writing. Given this vast database and the close links between a protein's structure and its function, much work has been devoted to classification of protein structure.



A The cartoon representation of the Cα backbone of 2-deoxyribose-5-phosphate aldolase. Secondary structure elements are coloured, pink are α-helices and yellow are β-strands.



B The atomic detailed representation of the Cα backbone of 2-deoxyribose-5-phosphate aldolase. Atoms are coloured according to their element symbol, eg water atoms are coloured in red.

Figure 1.1: Two different representations of 2-deoxyribose-5-phosphate aldolase.

The notion of protein structure can be divided into distinct but related levels. The sequence of amino acids that make up a protein is its **primary** structure, and is always identifiable. The start of the sequence of amino acids is known as the N-terminus, with the end known as the C-terminus. Researchers most often visualise the structure of a protein via a cartoon representation of its Cα backbone curve (as in Figure 1.1A), the discrete 3-dimensional curve made up of the central α carbon atom of each amino acid residue. The cartoon representation makes the specific local conformations of the backbone, known as the **secondary** structure, much more easily identifiable. We show the atomic detail representation of the same protein in Figure 1.1B, to highlight why the cartoon representation is preferred. The most common secondary structures are α helices and β strands, which can be identified by their hydrogen bonding patterns [2]. These secondary structures are both helical in nature and relatively uniform across all proteins. In this work, we will refer to subsections that do not fall into one of these two classes of secondary structure as linkers. Linkers can exhibit a much wider range of geometries across proteins. The variation of the linkers as they connect the other more rigid secondary structure elements leads to the complex entanglement of the backbone, known as

the **tertiary** structure. The arrangements of several adjacent secondary structures is referred to as **super secondary** structure, and will be one a key focus of this thesis alongside the tertiary structure.

1.2 Structure of the Thesis

The remainder of this chapter will be devoted to an overview of the existing work that motivates this thesis. This will include defining the protein folding problem and recent major breakthroughs in machine learning approaches to it. We will then introduce Biological Small Angle X-ray Scattering (BioSAXS), which is a technique for determining protein structures in their near native state in solution. We then present the constrained backbone algorithm, a technique for ab initio protein structure predictions from BioSAXS data. We will discuss some of its limitations, motivating our work on producing a bound on entanglement which restricts the search space. This work is one of my major contributions to the development of this algorithm into the open-source Carbonara software package as part of my PhD. We will conclude this chapter with a discussion of the various notions of structural similarity for proteins.

Chapter 2 will give a detailed background of the mathematical machinery we will use to address the goals of this thesis introduced in this chapter. We will define a discrete curvature and torsion, the two quantities which govern the constrained backbone algorithm, and discuss their relationship with other constraints on the secondary structure of proteins. We then give a more detailed discussion of the different approaches to measuring protein structural similarity, from alignment based methods to the more topological approach. Following this, we introduce the writhe, a measure of global self entanglement of a curve which is fundamental to the majority of the work in this thesis. We will discuss the existing applications of the writhe to protein studies, highlighting how the work within this thesis connects and builds upon this existing knowledge. We conclude Chapter 2 with a discussion of the concept of smoothing the backbone curve, and how it relates to applications of measures such as the writhe to protein structure.

Chapter 3 details a subset of the contents of my first publication [3], which introduced a novel method of backbone smoothing which produces a minimal representation of the global self entanglement of secondary structure elements. We show that this method of smoothing provides a natural definition of length for considering the scaling of complexity of entanglement for protein backbones. With this smoothing approach, we uncover two dominant super secondary motifs that are common across a diverse set of proteins.

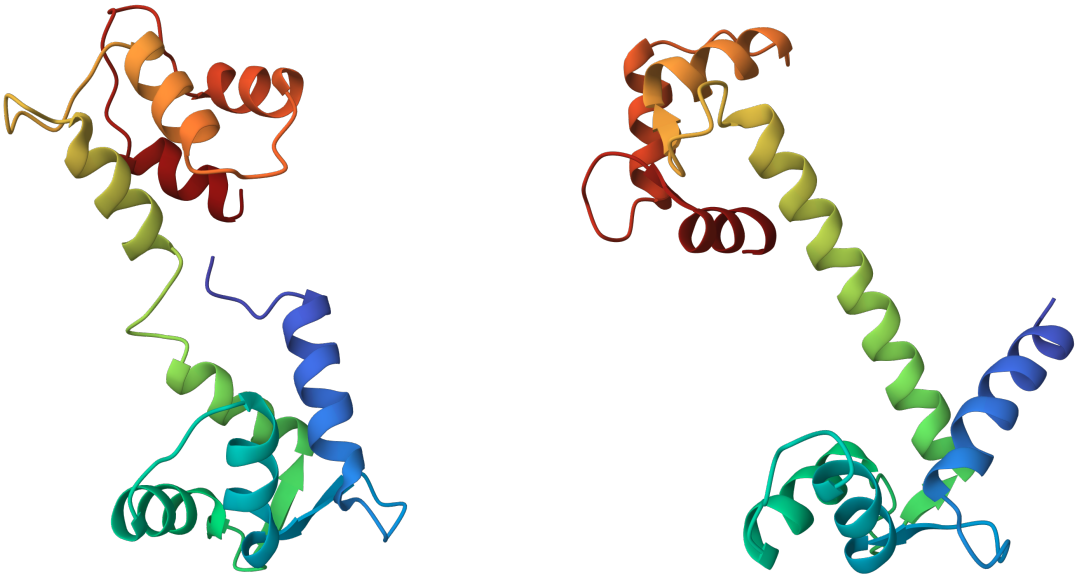
Chapter 4 begins with some further work from [3], deriving empirical bounds on the minimum absolute entanglement of a protein's backbone with respect to its secondary structure content. The remainder of this chapter is devoted to the developments of the Carbonara software package, with these empirical bounds included as a penalty during predictions to produce realistically entangled conformations. In this chapter we include an example of the way in which the development of Carbonara has shifted from ab initio prediction due to the prevalence of AlphaFold, highlighting where it can provide insight that AlphaFold cannot. We conclude this chapter by introducing the framework for a complementary model which could form part of a wider Carbonara prediction pipeline, providing a proof of concept of its application. The publication accompanying the release of Carbonara is still in preparation.

Chapter 5 builds upon the writhe similarity metric introduced in [3], taking advantage of both the novel backbone smoothing method and the bounds on entanglement presented in the preceding chapters. We provide some illustrative examples of the similarities detected by this metric, with an interactive iPython notebook available for researchers to benefit from these tools. We conclude this chapter by clustering our representative sample of the PDB using this similarity metric, revealing interesting relationships between the super secondary motifs.

1.3 Predicting Protein Structure

1.3.1 The Protein Folding Problem

One fundamental question that researchers try to address is can we predict the tertiary structure from the primary sequence alone. This is known as the protein folding



A An example Calmodulin structure with the disordered linker section (PDB: 1CFD [8])

B An example Calmodulin structure with the alpha-helical section (PDB: 1CLL [9])

Figure 1.2: Calmodulin, an example of a protein which can adopt multiple different tertiary structures

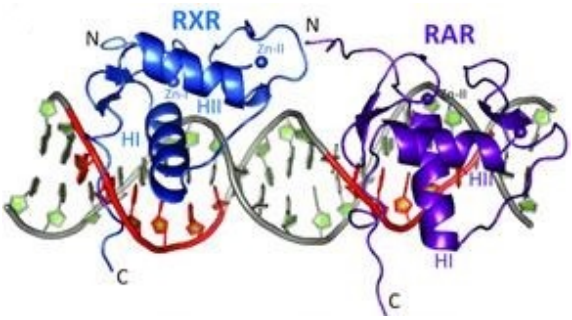
problem [4] and has seen much progress over the decades since its emergence in the 1960s. Perhaps the most significant recent advance was due to the introduction of machine learning (ML) approaches such as AlphaFold [5] and RoseTTAFold [6]. The power of these methods speaks for itself, with AlphaFold2 predicting tertiary structures for more than 200 million amino acid sequences. However, since these methods are trained on known structures from the PDB, we can expect low-confidence predictions for primary sequences for which there is no associated experimentally determined tertiary structure. Similarly, since these approaches are predictions based on sequence alone, they may have difficulty determining "the" structure of a protein which can adopt many conformations according to its environment. One particular class of proteins which present a challenge in this regard are Intrinsically Disorded Proteins (IDPs). These proteins contain one or more Intrinsically Disordered Regions (IDRs) which lack stable 3D structure under physiological conditions, instead adopting multiple differing conformations. For example, it was shown in [7] that AlphaFold is significantly outperformed by other models in terms of its accuracy and speed for prediciting disordered regions.

On the latter point, one such example is the calcium (Ca^{2+}) binding messenger protein Calmodulin. This protein has two roughly similar globular domains at the N and C termini, connected by a long, usually disordered linker section, as shown in Figure 1.2A. However, in a Ca^{2+} rich environment, this linker section can form a more ordered alpha-helical structure [10], as can be seen in Figure 1.2B. Since this variation in the tertiary structure is solely dependent on the protein's environment, it is an example of structural nuance which may be missed by structural predictions based on sequence alone. Though this specific interaction is well understood for Calmodulin, it provides a simple example of the limitations of taking a purely sequence based approach to predicting the tertiary structure.

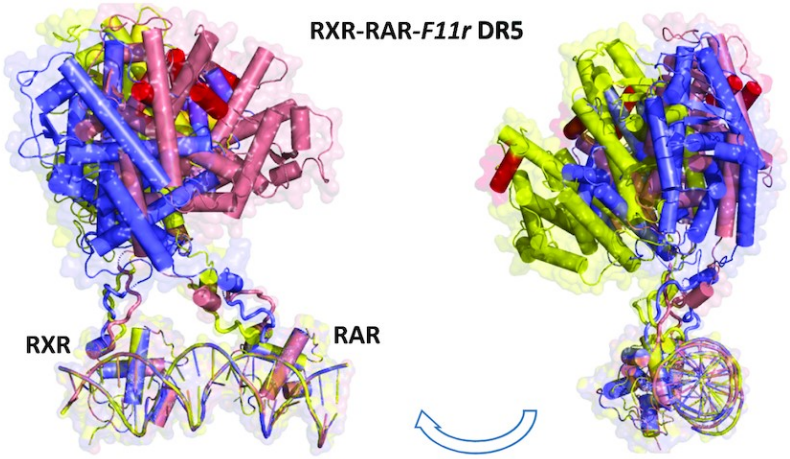
1.3.2 Biological Small Angle X-Ray Scattering

One technique that is particularly suited to situations where the inherent flexibility of the protein backbone may affect the predictions is Biological Small Angle X-Ray Scattering (BioSAXS). BioSAXS experiments allow researchers to study the size and shape of proteins under near-native-state conditions in solution. In this way, BioSAXS experiments are possible for proteins that may be inaccessible to other techniques such as X-ray crystallography. For example, in Figure 1.3A from [11] we see the crystal structure of the DNA binding domain of the Retinoic Acid Receptor (RAR) bound to DNA. RAR has multiple domains and forms heterodimers with the Retinoid X receptor in vivo. The regions between these domains are flexible, allowing for large motion and making the protein as a whole uncrystallisable. In Figure 1.3B, we see examples of these large motions which were determined through BioSAXs experiments, providing us with a better understanding of how RAR/RXR works.

It is particularly important to note here that structures predicted via Machine Learning methods such as AlphaFold, or even crystallographically-determined structures, will often fail to match the BioSAXS data [12]. For example, in [13] it was shown that AlphaFold fails to predict the holo form of a protein in $\tilde{3}0\%$ of cases. This highlights the importance of developing methods for reliably interpreting SAXS data in the absence of an accurate initial structural predictions. Despite the advantages



A Crystal structure of RXR-RAR DNA binding domain, with the DNA binding domains shown as ribbon diagrams in blue and violet and the DNA depicted with the sugar-phosphate backbone as grey and red ribbons, the base pairs as stick model.



B Overall structure of RXR-RAR-DNA complex. Three representative rigid-body models are shown in blue, pink, and yellow, with helices coloured in red.

Figure 1.3: Models of the RXR-RXR-DNA complex determined through BioSAXS experiments, reproduced from [11].

of studying proteins in near-native conditions in solution in this manner, there are some challenges. In particular, there is an unavoidable loss of information because of the random motion of proteins in solution. One fundamental assumption made when interpreting BioSAXs data is that the proteins essentially take up every possible orientation, which leads to an averaging of the detected scatter X-rays. As a result, only information about the protein's interatomic distances remains, not their spatial orientation. Many methods have been proposed to address this challenge of interpreting BioSAXS data. In [14] and its subsequent improved model [15], the protein is represented as a volume filled with densely packed spheres which can be fit to the scattering data. Similarly in [16] the protein is modeled as a chain of beads whose scattering can be calculated and optimized against experimental data. These so-called "bead" models allow for ab initio predictions, that is, predictions that do not require any initial conformation. However, since they lack any recognisable secondary structure, their outputs remain difficult to interpret [17].

Outside of ab-initio prediction, one other method of interpreting BioSAXS data is that of verification. That is, assuming an accurate 3D model of the structure is available, and computing its X-ray scattering curve to compare to the experimental data. One of the major advances in computing accurate scattering curves from a 3D model, first presented in [18], was the inclusion of the layer of water molecules which lie at the surface of the protein, and have a markedly different scattering effect to the bulk solution. This water layer is treated implicitly in [18] and [19], ie the individual water molecules are not modelled, whereas in [20] they are explicitly modelled using molecular dynamics.

Molecular Dynamics (MD) simulations generate atomic or near-atomic level detail on the movement of molecules. Further applications of MD to the verification of BioSAXS data have been coined SAXS-driven MD, introduced in [21] as part of the GROMACS software suite [22]. The atomic level detail they can provide comes with an obvious computational cost, which in turn means there is huge importance placed on the accuracy of the input configuration. This drawback will be particularly significant for a protein which may adopt multiple configurations in solution, such as that in Figure 1.2, where a significant search of the potential structure space

is required. An ab initio prediction method was introduced in [23] which is able to perform a rapid search of this structure space, thereby providing strong initial configuration for MD methods. We briefly discuss this approach in the following section, along with some of its limitations.

1.3.3 The Constrained Backbone Algorithm

The constrained backbone algorithm is so called as it represents the Cα backbone by a discrete curve whose local geometry is tightly constrained. By sampling new values for this local geometry, the algorithm produces new backbone models which better fit the experimental data, as shown in Figure 1.4. This approach is able to predict structures ab initio up to a $\tilde{7}.5$Å resolution, which, according to the Shannon sampling theorem described in [24], is the best resolution we can expect for most BioSAXS data. This method will provide the foundation for the development of the Carbonara software package in Chapter 4, where we will give a more thorough treatment of its mathematical background. For now, we will briefly discuss its limitations and how they motivate the work in this thesis.

By tightly constraining the local geometry of the backbone curve models, the method ensures that realistic secondary structure is preserved during predictions. However, this is not sufficient to guarantee that the super secondary and indeed tertiary structures produced will be realistic. Moreover, the fitting of a backbone model to an experimental scattering profile is under determined. That is, many different tertiary structures could produce a similar fit to the scattering data. To improve this model then, we would like to produce constraints on the super secondary and tertiary structure level which can restrict the search space to globally realistic conformations.

To achieve this, we need methods for categorising and quantifying the possible variation in super secondary and tertiary structures relative to the secondary structure. These methods should also be suitable for the resolution of scattering data, which is often limited to placing secondary structure elements in the correct relative position and orientation to form the tertiary structure. Moreover, with the prevalence of machine learning methods, we are often in a situation where we

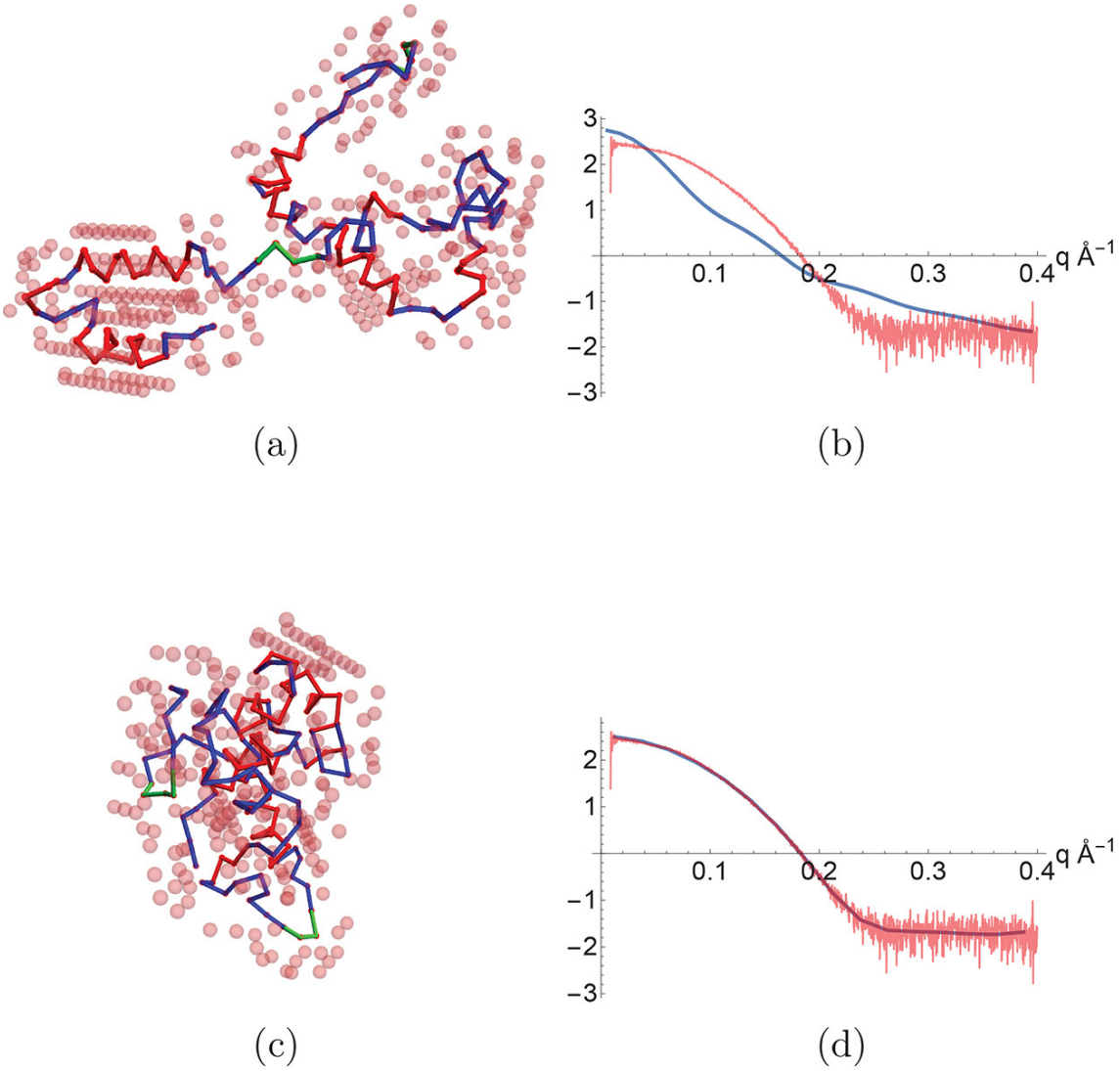


Figure 1.4: An illustration of the fitting procedure in the constrained backbone algorithm, taken from [23]. (a) An initial configuration, based on the secondary structure of Lysozyme. (b) The fit of the structure in (a) to the experimental scattering data. (c) The output configuration, which fits the experimental scattering data well as shown in (d)

already know the secondary structures but are trying to accurately predict their arrangement according to the experimental data. To understand the range of tertiary structures that are attainable for given secondary structure elements, we will first look to classify some common super secondary conformations.

1.4 Classifying Protein Structure and Flexibility

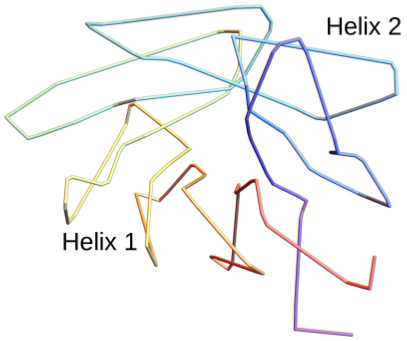
We now discuss one of the most common methods for categorising proteins based on their secondary structure content and tertiary structure. We will highlight through an example the limitations of this approach from our perspective.

1.4.1 Hierarchical Classification of Structure

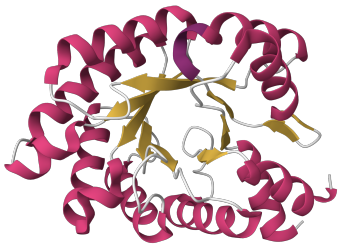
As discussed above, researchers now have millions of tertiary structure models at their disposal. The question of classifying and comparing these structures has therefore never been more vital. The CATH database [25] provides a hierarchical classification of structures using automated and manual methods. At the top level, Class (C), proteins are grouped according to their secondary structure content. Structures in the same Architecture (A) will have a high degree of structural similarity in the spatial arrangement of their secondary structures. On the Topology (T) level, the specific connection of secondary structure elements is included. At the finest level, Hierarchy (H), proteins are related provided there is sufficient evidence of evolutionary links between the domains. Similarly, the Structural Classification of Proteins Database (SCOP [26]) provides hierarchical classifications of proteins according to their structural and evolutionary relationships.

1.4.2 An example of the similarities missed by CATH

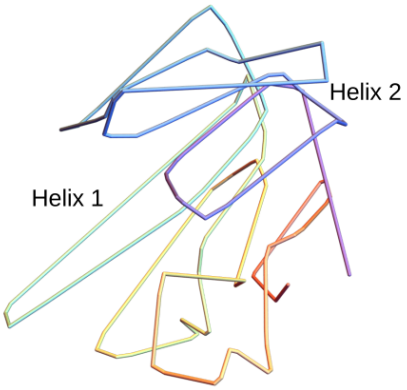
The classifications used in CATH are based on strict arrangements of secondary structure elements into domains. For example, the structure in Figure 1.5A is an example of a Tim Barrel, which is a topology under the Alpha-Beta architecture. This domain has by definition eight α helices and eight β strands that alternate along its backbone, forming the barrel-like structure. The location of the β strands



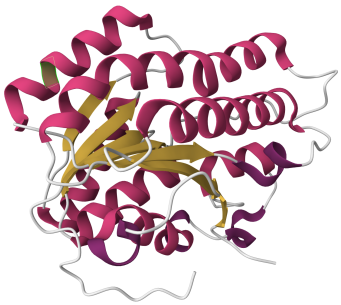
A The smoothed Cα backbone of the monomer unit of 2-deoxyribose-5-phosphate aldolase (PDB: 1P1X)



B The cartoon representation of the monomer unit of 2-deoxyribose-5-phosphate aldolase (PDB: 1P1X)



C The smoothed Cα backbone of an oxidoreductase, yciK from *E.coli* (PDB: 3F1L)



D The cartoon representation of an oxidoreductase, yciK from *E.coli* (PDB: 3F1L)

Figure 1.5: The Cα backbones of proteins belonging to two distinct CATH architectures, with secondary structure removed for visual clarity. Shared large scale helical domains are annotated.

on the inside of this barrel is a key factor in its stability [27]. The structure in Figure 1.5C is an example of a Rossmann fold, a topology in the 3-Layer (aba) sandwich architecture. A Rossmann fold is defined by its six β-strands, the first three of which are alternated by α helices, with this motif repeating for the latter three. These two definitions are clearly quite similar, despite these CATH families being separated on not just the Topology level, but also the broader Architecture level.

In fact, visual inspection of the two structures in Figure 1.5 indicates a shared super secondary and tertiary structure. The removal of the locally helical nature of the secondary structure elements reveals their arrangement into larger-scale helical domains. The relative locations and number of coils in these helices is common across the two structures. This specific example is more than just a curiosity. The results of [28] indicate a very close link between these two domains on the primary sequence, or even evolutionary level. In [28] the authors used directed evolution techniques in attempts to design a TIM barrel structure, but instead found that a Rossmann fold-like structure was produced. When these results were tested against the contemporary state-of-the-art computational techniques, none was able to predict the produced structure, and most agreed that the sequence should indeed produce a TIM barrel conformation.

In this thesis, the underlying motivation for classifying similar structures is to aid in constraining the state-space search for BioSAXS experiments. As such, our methods for classifying similarities should be robust to the resolution of data we would expect from a BioSAXS experiment. The previous example highlights the limitations of the CATH approach in this regard. In Chapter 2 we will introduce topological methods of classifying protein structures, which we show are better able to quantify similarities for low-resolution structures. In addition to constraining the state-space search, the metrics we develop allow for structural comparisons, which reveal surprising and interesting relationships between proteins. This forms another major part of my research in this thesis. In particular, we can investigate the relationship between the secondary structure and the range of possible tertiary structures, in some way similar to the hierarchical approach of CATH, where the

secondary structure content (Class) informs the tertiary structure (Architecture).

In fact, as discussed in [4], many of the theories of the mechanism by which proteins fold involve a hierarchical process. This may be the formation of secondary structures that then fold into the tertiary fold, as in [29], or it could be through the formation of the so-called microdomain structures [30] which diffuse and collide to form the tertiary structure. In either case, the range of possible folds for a protein's native state will be governed by the subunits of this hierarchical process. As we look to quantify this range of possible folds in this thesis, we will be considering the length scaling complexity of the entanglement of the protein backbone. Given this hierarchical process, we must pay careful attention to how we define the length of this relationship. In fact, we will see that the limits on entanglement of a protein backbone curve are functions of its secondary structure, as opposed to its number of amino acids.

1.5 Summary of the goals

We conclude this chapter by summarizing the goals of this thesis. In this thesis we will look to make precise the notions of similarity that are visually apparent in Figure 1.5. As discussed in [31], there is a theoretical limit on the accuracy of secondary structure predictions. Therefore, we should be careful when comparing and classifying protein structures according to strict rules regarding their secondary structure. Instead, we propose robust measures of structural similarity based on topological and global geometrical quantities that can identify the shared helical nature of the structures in Figure 1.5, as well as other interesting super secondary motifs.

This measure of similarity will be vital in addressing the search space problem of the constrained backbone algorithm. The resolution of BioSAXS data is often poor due to the random motion of proteins in solution. Any notion of similarity between predicted structures should therefore be robust to small-scale variations, whilst being sensitive to large-scale conformational changes. Structural similarity measures derived from topological and geometrical quantities are uniquely suited to

this problem. Using empirical limits on the range of possible tertiary structures determined by this similarity measure, we aim to reduce the search space for BioSAXS predictions to realistically folded structures.

CHAPTER 2

Background

In this chapter we give a detailed introduction of the geometrical and topological tools we will use to address the goals of the thesis presented in Chapter 1. We will first introduce the curvature and torsion, which are geometric quantities that are heavily constrained within secondary structures. We then discuss some of the existing methods for comparing tertiary structures. Following this, we introduce the writhe, a measure of global self-entanglement of a curve, and justify its use as the foundation for the structural similarity measures we use in this thesis. We also detail the areas of the literature that the work in this thesis seeks to build on.

2.1 Parameterising secondary structure

2.1.1 Ramachandran angles

The constrained backbone algorithm introduced in Chapter 1 is so called due to the local constraints on the geometry of secondary structure elements. Classically, the local geometry of the protein backbone is represented by Ramachandran plots. These plots are a method of visualising the energetically favourable dihedral backbone angles for amino acids [32]. The dihedral angles are the angles between the

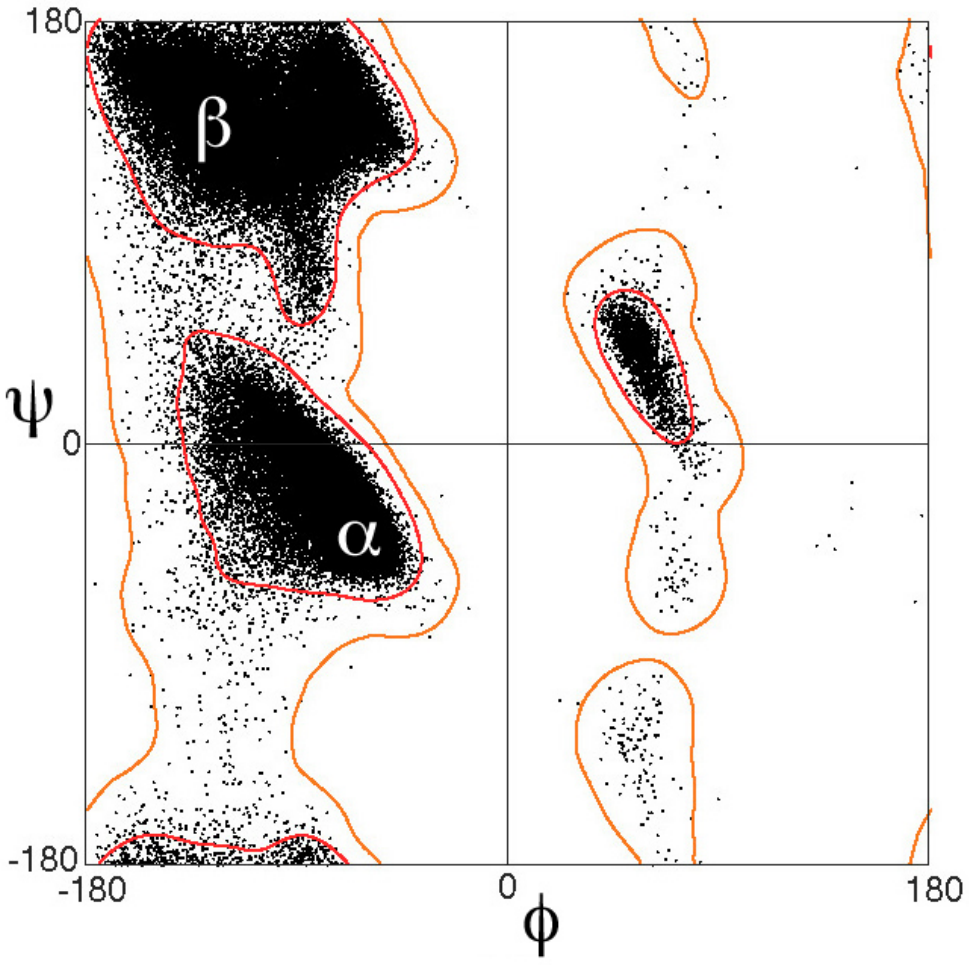
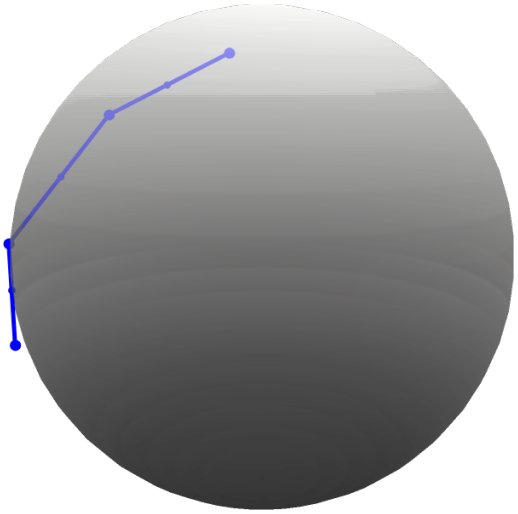
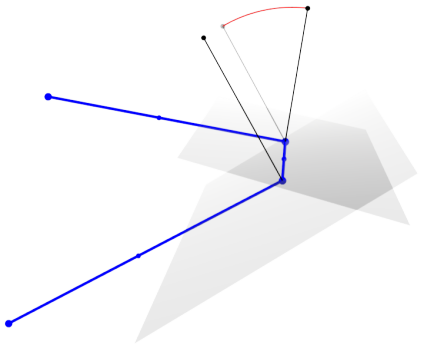


Figure 2.1: An example of a Ramachandran plot with data from high resolution crystal structures. The regions corresponding to secondary structure elements are highlighted. From `https://proteopedia.org/wiki/index.php/Tutorial:Ramachandran_principle_and_phi_psi_angles`



A In blue, a subsection of a discrete curve with four points. In grey, the inscribed sphere defined by the three midpoints of the curve. The curvature is the inverse of the radius of this sphere.



B In blue, a subsection of a discrete curve with four points. In grey, the two planes defined the first and last three points of the curve. In solid black, the normals to these planes. The torsion is the angle between these plane normals, as shown by the shadow of the translated first plane normal, and the red arc.

Figure 2.2: Visual representation of the discrete curvature and torsion quantities.

planes defined by the central Cα atom and its neighbouring carbonyl carbon (C) and amide nitrogen (N) atoms. An example of a Ramachandran plot is shown in Figure 2.1, note, in particular, that secondary structure elements occupy a tightly constrained region of the full potential space of angles. Ramachandran plots can also be used as a method of structural validation, by comparing the empirical distribution of dihedral backbone angles for a proposed structure against these favourible regions.

2.1.2 Defining discrete curvature and torsion

In the constrained backbone algorithm, we are modelling the discrete curve whose points are the Cα atoms. Without the other atoms present, the Ramachandran angles as defined cannot be computed, and therefore we require analogous geometrical quantities defined solely on the Cα backbone. For this, we use **curvature** κ and **torsion** τ. These are fundamental quantities of the local geometry of smooth space curves, which we will use to inspire a definition of discrete analogues.

Let $\boldsymbol{\gamma}$ be a smooth curve parameterised by t such that points on the curve are given by $\boldsymbol{\gamma}(t)$. Let $\boldsymbol{\gamma}'$ and $\boldsymbol{\gamma}''$ denote the first and second derivatives of $\boldsymbol{\gamma}$ with respect

to t respectively. The curvature of $\boldsymbol{\gamma}$ is given by

$$\kappa(t) = \frac{\|\boldsymbol{\gamma}'(t) \times \boldsymbol{\gamma}''(t)\|}{\|\boldsymbol{\gamma}''(t)\|^3}. \tag{2.1}$$

The curvature is so-called as it measures the deviation of a curve from being a straight line. In this expression for the curvature, we measure the rate of change of $\frac{\boldsymbol{\gamma}'(t)}{\|\boldsymbol{\gamma}(t)\|}$ with respect to arc length. If $\boldsymbol{\gamma}(t)$ was parameterised by arc length, the curvature would simply be given by $\|\boldsymbol{\gamma}''(t)\|$. The torsion of $\boldsymbol{\gamma}$ is given by

$$\tau(t) = \frac{(\boldsymbol{\gamma}'(t) \times \boldsymbol{\gamma}''(t)) \cdot \boldsymbol{\gamma}'''(t)}{\|\boldsymbol{\gamma}'(t) \times \boldsymbol{\gamma}''(t)\|^2}. \tag{2.2}$$

The torsion measures the deviation of the curve from its plane. The curvature and torsion together are sufficient to uniquely define a smooth curve up to rotation and translation (this is coined the Fundamental Theorem of Space Curves on p.84 of [33]). Given we are interested in the discrete backbone curve defined by the central Cα atom of the amino acids, we look for discrete analogues to these quantities to define our curve. We define a discrete curvature as follows:

$$\kappa = \frac{2|\sin(\theta)|}{\|\boldsymbol{x}_{m_i} - \boldsymbol{x}_{m_{i+1}}\|}, \tag{2.3}$$

where $\boldsymbol{x}_{m_i}$ is the midpoint of $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$, and θ is the angle between the vectors $\boldsymbol{x}_{m_i} - \boldsymbol{x}_{m_{i+2}}$ and $\boldsymbol{x}_{m_{i+1}} - \boldsymbol{x}_{m_{i+2}}$. The discrete curvature is the inverse of the radius of the inscribed sphere uniquely defined by the three midpoints, as seen in Figure 2.2.

Given four points, we can define two planes with unit normal vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ by

$$\begin{aligned} \boldsymbol{n}_1 &= \frac{(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) \times (\boldsymbol{x}_{i+2} - \boldsymbol{x}_{i+1})}{\|(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i) \times (\boldsymbol{x}_{i+2} - \boldsymbol{x}_{i+1})\|}, \\ \boldsymbol{n}_2 &= \frac{(\boldsymbol{x}_{i+2} - \boldsymbol{x}_{i+1}) \times (\boldsymbol{x}_{i+3} - \boldsymbol{x}_{i+2})}{\|(\boldsymbol{x}_{i+2} - \boldsymbol{x}_{i+1}) \times (\boldsymbol{x}_{i+3} - \boldsymbol{x}_{i+2})\|}. \end{aligned} \tag{2.4}$$

Torsion is given by:

$$\tau = \frac{2\sin(\varphi/2)}{\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\| + \|\boldsymbol{x}_{i+2} - \boldsymbol{x}_{i+1}\| + \|\boldsymbol{x}_{i+3} - \boldsymbol{x}_{i+2}\|}, \tag{2.5}$$

where φ is here the angle between the vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$. The torsion therefore measures the length-weighted angle between the two plane normals, as shown in Figure 2.2. It can be thought of as the tendency for the curve to leave the plane, so a circle has zero torsion but a helix has constant non-zero torsion.

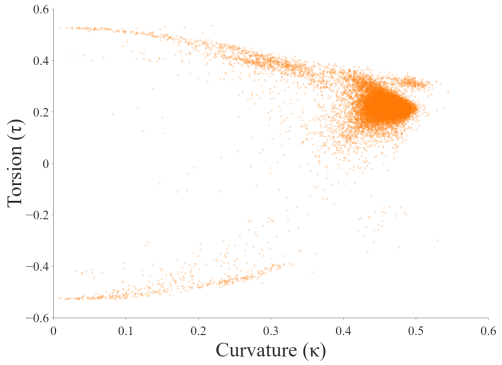
In order to measure the angle between the two planes, we must define a direction about which to rotate the first plane to be planar to the second. For this, we imagine translating $\boldsymbol{n}_1$ along the edge $\boldsymbol{x}_{i+2} - \boldsymbol{x}_{i+1}$ since this edge is shared by the two planes. We therefore define the rotation to be about this edge, giving the sign of θ. This is represented by the red arc in Figure 2.2B.

2.1.3 Curvature-Torsion space

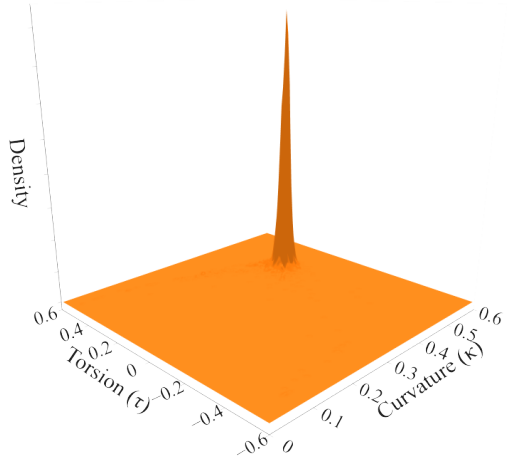
To investigate the distribution of these quantities across the PDB, we construct a representative sample of proteins from the PDB via the following criteria:

1. Good resolution (<2Å)
2. Consisting of between 30-300 residues.
3. Redundancies removed at 70% sequence identity.
4. Good model quality: $R_{work} \in [0, 0.2]$ and $R_{free} \in [0, 0.25]$.

On the third point, the PDB contains many entries consisting of the same protein sequence, studied under different experimental conditions or from different species. To avoid biasing our results towards proteins with multiple entries, we select one representative for all proteins which share over 70% sequence similarity. The crystallographic model quality parameter R_{work} measures the agreement of the calculated diffraction data for the structural model with the experimentally observed diffraction. R_{free} is computed similarly, but for a randomly selected subset of the experimental data, to avoid overfitting of the model. These parameters, combined with the resolution filter, ensure our representative sample of the PDB containts high quality structures to study. This yielded 10736 entries from the PDB at the time of query. For any entry that was made up of more than one chain, we only consider the first chain in the PDB file. We will use this same sample throughout this



A A scatter plot of the distribution of curvature and torsion for α-helices.



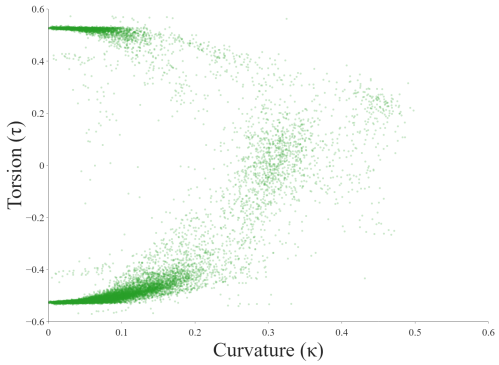
B A surface plot of the distribution of curvature and torsion for α-helices.

Figure 2.3: The distribution of curvature and torsion for α-helices from a representative sample of the PDB.

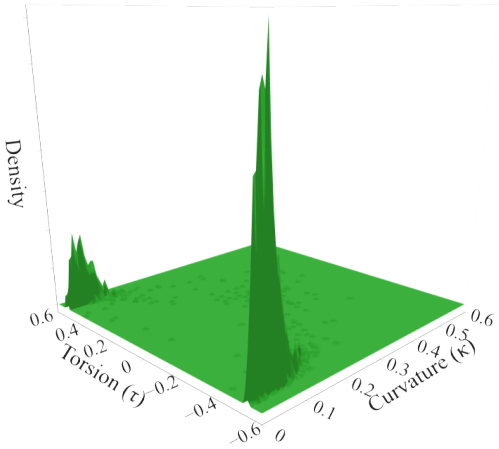
thesis. In Figure 2.3 we see that α helices have high curvature and positive torsion because of their consistent right-handed helical shape. β-strands are also helical, but with a much less tight coiling than the α-helices. They also differ from α-helices in that they can have both right- and left-handed coiling. This is represented in Figure 2.4 by the symmetric ridge-like regions of low curvature and high torsion, both positive and negative. The presence of these regions in Figure 2.3A is due to poor secondary structure labeling as opposed to being representative of possible geometries for α-helices, as seen in Figure 2.3B. Linker subsections have a much more uniform distribution of geometries, sharing similar peaks with α -helices and β -strands, as seen in Figure 2.5. This variety in local geometry is vital for linkers in their role in connecting the other more rigid secondary structures and allowing the formation of the global fold. On the other hand, we can think of α-helices and β-strands acting as barriers to entanglement on a global scale, an idea that will be developed much further in Chapter 3.

2.1.4 Limitations of the curvature-torsion parameterisation

To generate new backbone curves, the model samples values from these distributions to construct a new curve subsection. In particular, we consider three initial points

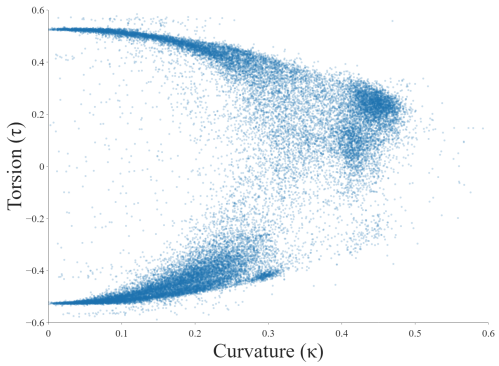


A A scatter plot of the distribution of curvature and torsion for β-strands.

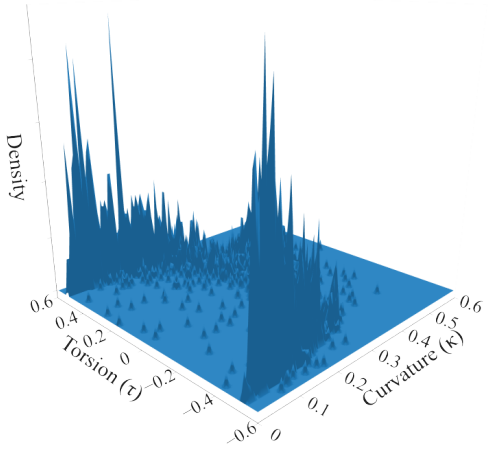


B A surface plot of the distribution of curvature and torsion for β-strands.

Figure 2.4: The distribution of curvature and torsion for β-strands from a representative sample of the PDB.



A A scatter plot of the distribution of curvature and torsion for linkers.



B A surface plot of the distribution of curvature and torsion for linkers.

Figure 2.5: The distribution of curvature and torsion for linkers from a representative sample of the PDB.

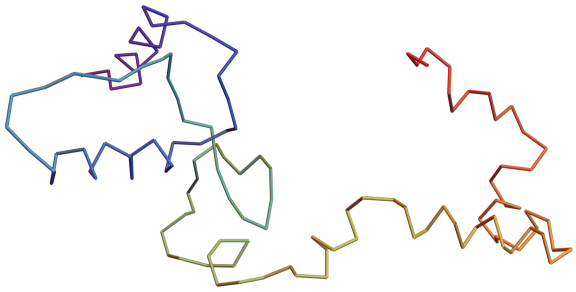


Figure 2.6: An example of constrained backbone model whose global geometry is unrealistic.

$\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ with fixed separation $R = 3.8$. Up to rotation and translation, we can freely chose the positions of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ with this separation, then $\boldsymbol{x}_3$ is placed such that the distance between itself and $\boldsymbol{x}_1$ is greater than R. The next point in the curve will be given by

$$\boldsymbol{x}_4 = \boldsymbol{x}_3 + R(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta), \tag{2.6}$$

for some $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi]$. From the four points defined by $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \theta, \phi)$ we can compute a curvature and torsion. Conversely, sampling a value for κ and τ, we can solve Equations (2.3) and (2.5) for θ and ϕ giving $\boldsymbol{x}_4$. By sampling from these heavily constrained distributions, the model ensures that the local geometry of the produced backbone curve is realistic. In fact, it is shown in the supplement of [23] that the preferred regions of the curvature-torsion space correspond to those of the Ramachandran space. The curvature-torsion angles here are sampled independently for each new point in the backbone curve. Though there may be some correlation between subsequent curvature-torsion pairs within secondary structure elements, the model at present does not exploit this. Primarily this is to keep the computationaly complexity of producing new conformations as minimal as possible, to allow for rapid search of the conformational landscape.

Sampling from these local geometry distributions does not necessarily imply that the backbone will have a realistic structure globally. For example, in Figure 2.6 we show an example of a backbone curve with the same secondary structure as the Calmodulin structure shown in Figure 1.2A. This conformation how is not realistic, with the secondary structures not sufficiently tightly packed to be stable. We there-

fore require methods for comparing and classifying predicted structures according to their tertiary structure, which we will discuss in the following section.

To conclude this subsection, we note some of the other backbone curve modeling approaches taken and justify the use of the constrained backbone algorithm. In [34], Hausrath et al. use a continuous approach based on fitting polyhelices with given curvature profiles. This method is shown to be effective for α-helically dominated backbones, however due to the helical nature of its definition, the model has limited accuracy for β-sheets and linkers. A soliton representation is proposed in [35], where it is used to build a library of 200 sets of loop parameters, which when combined α-helices and β-strands, can be pieced together to produce a backbone model. Krokhotin et al. show that this model can effectively reproduce the majority of the high resolution structures in their dataset. Despite this method's impressive results, we are looking for a model with a much smaller number of parameters for ease of manipulation with respect to the BioSAXS data. The motivation for the specific approach taken in the constrained backbone algorithm comes from the fact that curvature and torsion are sufficient to uniquely define a smooth curve. By finding discrete analogues, we can characterise discrete curves with just these two parameters.

2.2 Comparing tertiary structure

Now that we have established the tools we will use to study the local structure of proteins, we move to the global structure. The existing methods for global structure comparison can broadly be put into two categories. The first, alignment-based methods, require the two backbones to be aligned in some way before computing some pairwise distance score. We will show that these methods have specific limitations that make them unsuitable for our approach. We then introduce topological methods of structure comparison that can address these limitations. We then introduce the writhe and justify its use as a similarity metric over some of the alternative topological methods we discuss.

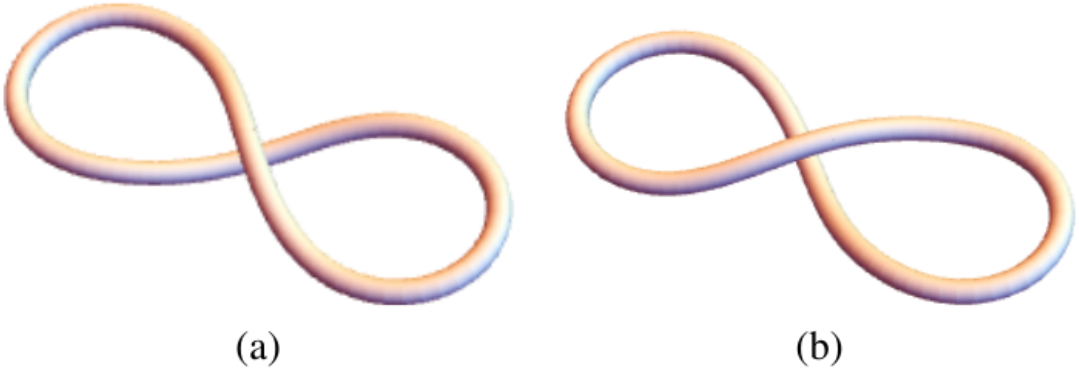


Figure 2.7: Two different figure 8 loops.

2.2.1 Alignment based methods

The most commonly used methods for protein structural comparison are based on root mean square comparisons [36]. Two structures are rotated such that the sum of the minimum Euclidean distance between sequentially aligned points is minimised, the so-called root mean squared distance (RMSD). Very often the comparison is between similar subsections, for example one can compare hinging proteins this way. The DALI method [37] is an alignment-based method that seeks the largest common substructure between the two proteins. TM-align [38] uses an RMSD-based metric which is less sensitive to local variations than pure RMSD, called TM-score [39]. The jFATCAT method [40] uses a more flexible alignment approach, allowing a certain number of twists in the backbone to align the residues. However, alignment-based distance metrics, despite their many merits, are not suitable for the types of similarity that we are aiming to quantify in this thesis.

Consider, for example, the two figure 8 curves shown in Figure 2.7. These curves differ only by the relative orientation of the crossing. The RMSD between these two conformations, as measured by discretisation of the curves seen in Figure 2.7, would be very small. This is because one could move from one conformation to the other by moving only a very small subset of points. However, this change would not be accessible to realistic protein dynamics; it would require cutting the curve or passing it through itself (analogous to the breaking of covalent peptide bonds in the backbone). In order to represent this fundamental difference between these

two curves, we must use a similarity measure which is sensitive to features such as crossings. For this, we turn to topological metrics.

2.2.2 Introducing topological methods

Topological methods are those derived from aspects of Knot theory [41,42], of which our method can be said to belong. Topological metrics are invariant to rotations and translations and do not require alignment. A second potential advantage is that they classify structures up to isotopy, that is, they measure two structures as similar if they can be distorted into each other without having to construct the distortion. As shown in the supporting information of [23], metrics derived from topological quantities can classify the two figure 8 curves in Figure 2.7 as significantly different. It is important to note here that despite this advantage, direct application of topological metrics may not distinguish between two differently folded proteins who global topology is the same. It for this reason researchers have taken to measuring the topology of substructures within proteins to highlight their difference, a technique we will build upon.

2.2.3 Knot theory approach to protein structure classification

The most common topological measures that are applied to protein structures come from the world of knot theory. Mathematically, a knot is an embedding of a circle in a three-dimensional space. However, proteins have distinct end points and, therefore, are not mathematical knots. The first example of bridging this gap between knot theory and proteins was in 1994, where Mansfield proposed that we "close" a protein by projecting its endpoints to a point on a sphere surrounding the backbone curve [41]. To this closed curve we can assign a knot type. The choice of projection will affect the knot type; therefore, we often imagine many projections to determine the most common knot type [43]. This initial survey of knottedness in the PDB found just one example of a knotted protein backbone amongst the 400 entries, indicating that though mathematically interesting, knotting in proteins is rare.

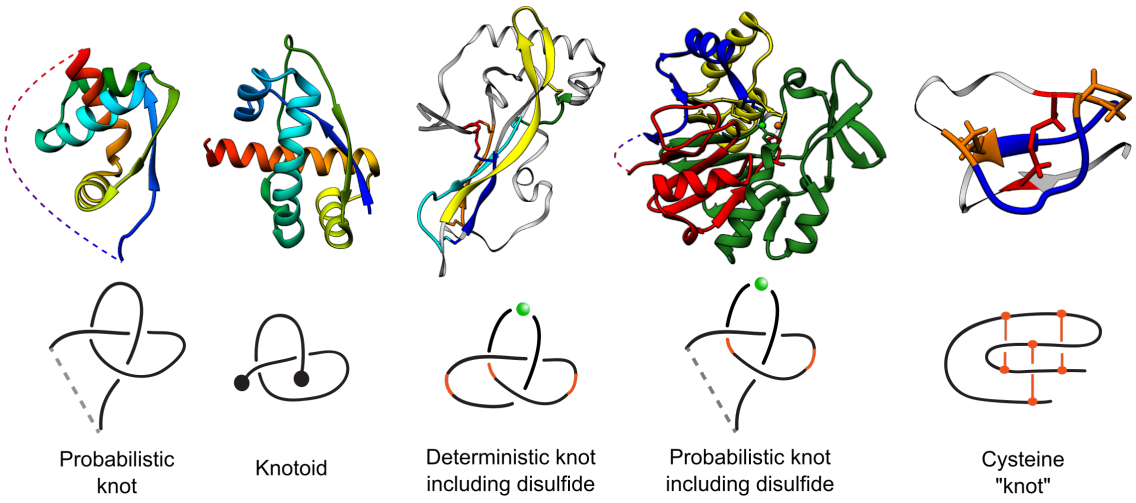


Figure 2.8: The various types of knotting that are identified by KnotProt 2.0, from [45].

This method of detecting knottedness in proteins was further developed in 2000 by Taylor [44]. This work proposes a chain smoothing technique which leaves the end points fixed and removes intermediate points, provided that there is no change to the initial entanglement of the curve with this removal. In this way, the protein is reduced to a minimal representation of its global entanglement. For example, this smoothing method would transform an unknotted backbone curve into simply a straight edge connecting its endpoints after sufficient iterations. For those proteins which were not reduced to just two points, their endpoints would be safely away from the rest of the curve such that the probabilistic closure of Mansfield is no longer necessary. This idea of smoothing the backbone to uncover entanglement will be developed further in chapter 3

The rarity of knotted protein backbones led King et al. to coin the term slipknots in [46]. A slipknot is defined as a conformation in which a subsection of the chain is knotted but the whole chain is classified as unknotted. The most common example of this involves the chain turning back on itself at some point to effectively "untie" the knot formed by the rest of the chain. With this definition, the KnotProt database was formed in [47], containing all currently known instances of (slip)knotting in proteins. This database was further developed in [45] to include entanglements in terms of knotoids (which we will introduce shortly), as well as entanglements due to interaction of ions or disulfide bonds as shown in Figure 2.8.

Introduced by Turaev in 2012, knotoids [48] provide an alternative extension to the knot-theoretic approach to protein structure. Knotoids are defined on line segments as opposed to circles, with two knotoids being equivalent if they differ only by continuous deformations away form their endpoints. Thus, they are a much more natural object to consider when studying protein chains. It was shown in [49] that the knotoid approach allows us distinguish proteins that we could not tell apart using the classical knot theory techniques. Using a knotoid based approach to the protein folding problem, the Barbensi et al. in [50] were able to detect a novel folding pathway for the shallow knotted Carbonic Anhydrases.

2.3 The writhe and average crossing number

Despite the extensions to the theory discussed in the previous section, the problem remains that 99% of proteins, although they have complex entangled structures, do not contain any of the various forms of knotting [45]. As a result, researchers have turned to writhe-based measures as a method to study proteins. This is because the writhe is a measure of self-entanglement of a curve that is sensitive to topological features such as crossings as well as global geometric notions such as helical coiling.

We define a smooth curve $\boldsymbol{\gamma}$, parameterised by t such that points on the curve are given by $\boldsymbol{\gamma}(t)$. The tangent vector to this curve is denoted $\boldsymbol{T}(t) = \frac{\mathrm{d}\boldsymbol{\gamma}}{\mathrm{d}t}$. The writhe of the curve $\boldsymbol{\gamma}$ is given by the Gauss linking integral [51]

$$Wr = \frac{1}{4\pi} \int_{\gamma} \int_{\gamma} \boldsymbol{T}(s) \times \boldsymbol{T}(t) \cdot \frac{\boldsymbol{\gamma}(s) - \boldsymbol{\gamma}(t)}{\|\boldsymbol{\gamma}(s) - \boldsymbol{\gamma}(t)\|^3} \, \mathrm{d}s \, \mathrm{d}t. \tag{2.7}$$

A heuristic explanation of this formula is as follows. We first assign an orientation to $\boldsymbol{\gamma}$, essentially deciding a direction of travel along the curve. We then consider viewing $\boldsymbol{\gamma}$ embedded in space from some point away from the curve. There may be points where the curve passes over itself, which we call crossings. As a standard convention, a crossing is called positive if the strand passing underneath travels from right to left and negative if it passes left to right, according to the assigned orientation, as demonstrated in Figure 2.9. The number and sign of these crossings will depend on the viewpoint. The double integral above then essentially averages

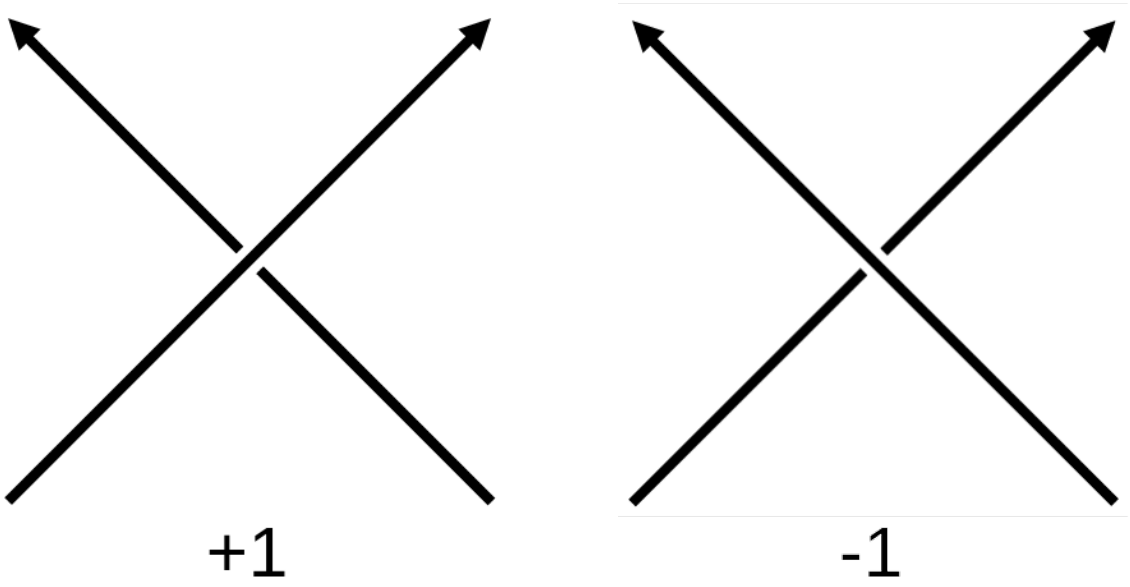


Figure 2.9: The crossing sign convention for orientated curves.

the signed sum of the crossings seen by observers from all possible points of view [52].

In this thesis, we focus on the discrete backbone curve, so we use the discrete analogue of the writhe of [53]. The writhe of a discrete curve $\mathcal{C}$ of length n is given by

$$Wr(\mathcal{C}) = 2\sum_{i=2}^{n-1}\sum_{j<i}\frac{\Omega_{ij}}{4\pi}, \tag{2.8}$$

where Ω_{ij} represents the contribution to eq. (2.7) from the crossing of edges connecting $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+1}$ and $\boldsymbol{x}_j$ to $\boldsymbol{x}_{j+1}$. There are numerous equivalent methods for computing Ω, and we will follow Method 1a given in [53]. For this, we denote by $\boldsymbol{r}_{i,j}$ the edge between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We then define the unit normal vectors:

$$\boldsymbol{n}_1 = \frac{\boldsymbol{r}_{i,j} \times \boldsymbol{r}_{i,j+1}}{\|\boldsymbol{r}_{i,j} \times \boldsymbol{r}_{i,j+1}\|}, \tag{2.9}$$

$$\boldsymbol{n}_2 = \frac{\boldsymbol{r}_{i,j+1} \times \boldsymbol{r}_{i+1,j+1}}{\|\boldsymbol{r}_{i,j+1} \times \boldsymbol{r}_{i+1,j+1}\|}, \tag{2.10}$$

$$\boldsymbol{n}_3 = \frac{\boldsymbol{r}_{i+1,j+1} \times \boldsymbol{r}_{i+1,j}}{\|\boldsymbol{r}_{i+1,j+1} \times \boldsymbol{r}_{i+1,j}\|}, \tag{2.11}$$

$$\boldsymbol{n}_4 = \frac{\boldsymbol{r}_{i+1,j} \times \boldsymbol{r}_{i,j}}{\|\boldsymbol{r}_{i+1,j} \times \boldsymbol{r}_{i,j}\|} \tag{2.12}$$

We consider the sum of the angles between these vectors

$$\Omega^* = \sin^{-1}(\boldsymbol{n}_1 \cdot \boldsymbol{n}_2) + \sin^{-1}(\boldsymbol{n}_2 \cdot \boldsymbol{n}_3) + \sin^{-1}(\boldsymbol{n}_3 \cdot \boldsymbol{n}_4) + \sin^{-1}(\boldsymbol{n}_4 \cdot \boldsymbol{n}_1). \quad (2.13)$$

Then the evaluation of the Gauss integral from the crossing of $\boldsymbol{r}_{i,i+1}$ and $\boldsymbol{r}_{j,j+1}$ is given by

$$\frac{\Omega_{ij}}{4\pi} = \frac{\Omega^*}{4\pi} \operatorname{sgn}((\boldsymbol{r}_{j,j+1} \times \boldsymbol{r}_{i,i+1})\boldsymbol{r}_{i,j}) \quad (2.14)$$

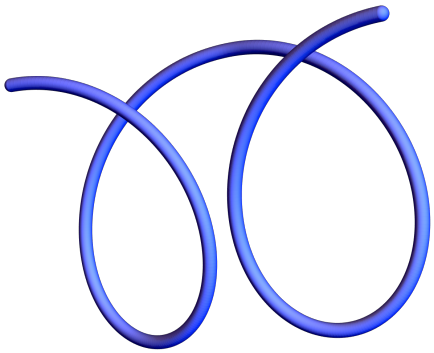
In chapter 4 we look to derive bounds on entanglement for protein backbone curves. For this, it will be useful to have a positive definite measure of the complexity of the fold. We use the average crossing number (acn), given by

$$acn(\mathcal{C}) = 2\sum_{i=2}^{n-1}\sum_{j<i}\frac{|\Omega_{ij}|}{4\pi}, \quad (2.15)$$

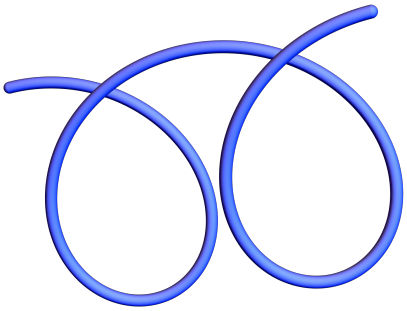
which is a variation of the writhe which ignores the sign of the crossings in the sum. In Figure 2.10A we see a curve that has two crossings with opposite sign. As a result, the writhe of this curve is close to zero at -0.00156. On the other hand, the acn of this curve is 1.69. The crossings of the curve in Figure 2.10B have the same sign. This is reflected in the writhe being of similar magnitude to the acn at -1.51 and 1.91 respectively. This simple example highlights the different information that these two quantities capture. Finally in Figure 2.10C we show an example of a trefoil knot. The writhe of this curve is 3.35, reflecting the defining topology of this knot with its three crossings.

2.3.1 Applications of the writhe to proteins

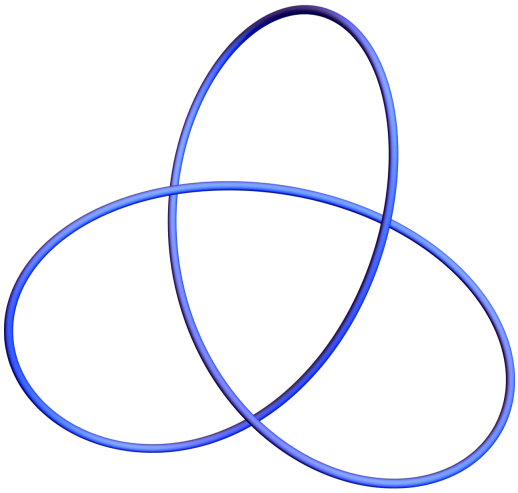
Applications of the writhe as a measure of self-entanglement are already well established in DNA research. For example, in [54] Fosado et al. measure the relaxation of writhe as it relates to the dynamics of double-stranded DNA supercoiling. In [55] a fundamental theorem of the conservation of writhe plus its related quantity "twist"" for supercoiled circular DNA is exploited to tune DNA mobility. For a more comprehensive review of advances in this area, see [56]. The first application of writhe in the context of proteins is in [57] where it is used as a verification of the correct



A A curve with two crossings of opposite sign.



B A curve with two crossings of the same sign.



C A trefoil curve with three crossings.

Figure 2.10: Two simple curves which highlight the different information captured by the writhe and *acn*.

terminus threading of the bovine pancreatic trypsin inhibitor protein in simulation. This work took advantage of a fundamental property of the writhe, namely that its value jumps by ± 2 when a curve passes through itself. With this property, conformations whose writhe differs from the native X-ray conformation by 2 can be identified as having an incorrect terminus threading.

The writhe was also used in [58] to study not just the tertiary structure of proteins, but also the scaling of entanglement as it relates to the secondary structure. This builds on the idea in [59] that secondary structures essentially act as rigid edges when viewed as part of the global entanglement. Therefore, for two proteins with a similar number of residues, we could expect the protein that has fewer secondary structure elements to be less entangled, as measured by the writhe. This relationship between the scaling of the entanglement and the secondary structure will be further discussed in chapter 3

Perhaps the most notable contributions to the protein writhe literature were made by Røgen and Bohr in [60]. Considering generalizations of the Gauss integral, a 30-dimensional vector is formed of the scores for a given protein on these measures, which has been shown to agree well with the CATH classification in [61]. In [62] this idea was further developed, with a comparison metric based on the writhe value of fixed length protein subsets validated against the SCOP classification, again with good results. In [63] a more rapid metric was created from the same Gaussian integrals as in the other two studies, and it was once again shown to compare favourably with the SCOP benchmark set. The Gaussian-integral approach was also applied to subsets (fragments) of the protein, through a fingerprinting technique, using the entanglement of subchains to identify rare conformations in proteins in [64]. Despite the evident success of these methods, there is difficulty in interpreting these higher order Gauss integrals intuitively.

2.3.2 Defining writhe fingerprints and writhe profiles

In Chapter 3 we will use a similar fingerprinting technique to identify large-scale helical geometries present in various families of proteins which have an important role in their stability. We define the **writhe fingerprint** as follows. We will consider all

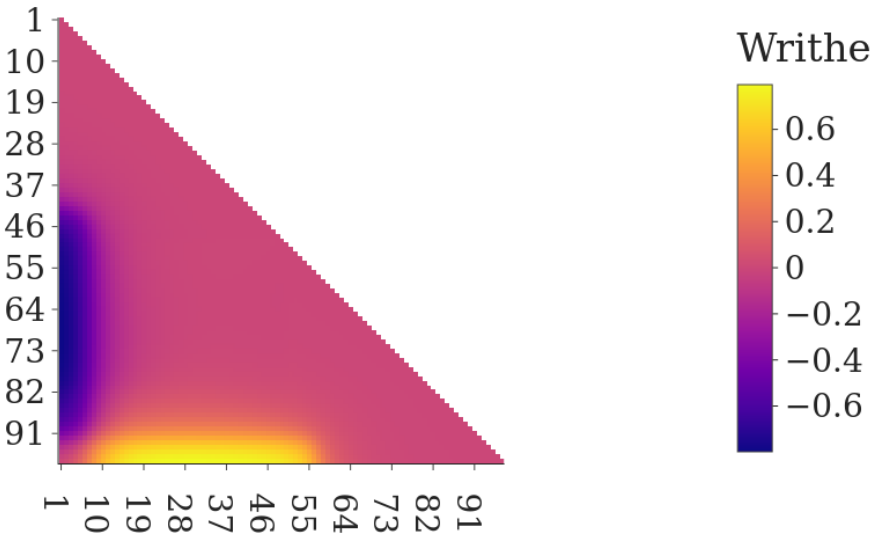


Figure 2.11: An example of a writhe fingerprint for the curve shown in Figure 2.10A

sub-curves of length at least four, since we need at least four points for a meaningful writhe calculation (four points define the two edges whose mutual entanglement we calculate in Ω_{ij} of Equation (2.8)). We will denote these subcurves $\mathcal{C}_{(i,j)}$ where i and j are the start and end indices respectively (so $\mathcal{C}_{(1,n)}$ corresponds to the full curve). This gives rise to a $(n-4) \times (n-4)$ lower triangular matrix whose $(i,j)^{\text{th}}$ entry is $Wr(\mathcal{C}_{(i,j+4)})$. In [65], the Ω_{ij} are called the local segment-to-segment (StS) writhe, and are shown to be useful geometric signatures as part of the training of a Neural Network to detect topological invariants and knot polynomials.

In Figure 2.11 we give a visual example of a writhe fingerprint for the curve seen in Figure 2.10A. There are two clear regions where the writhe peaks in magnitude but with opposing sign, corresponding to opposing orientation of the two crossings in the curve. This simple example highlights the type of information that we can extract from the writhe fingerprint. Another feature of the writhe fingerprint we can look at is the variation along specific rows/columns. For instance, the first column contains $Wr(\mathcal{C}_{(1,k)})$ for $k = 5, \ldots, n$, which allows us to study the change in writhe as we include more of the curve. In what follows, we will refer to the graph of $Wr(\mathcal{C}_{(1,k)})$ for $k = 5, \ldots, n$ as the **writhe profile** of $\mathcal{C}$. We define the *acn* fingerprint and *acn* profile analogously.

Using fundamental properties of the writhe, we will look to uncover groups of

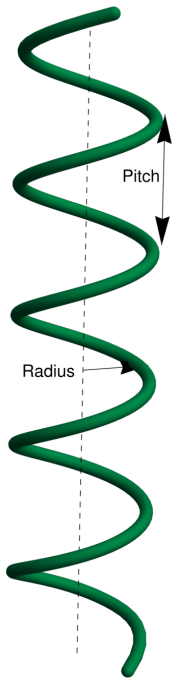


Figure 2.12: A helix, with the pitch P and radius R annotated.

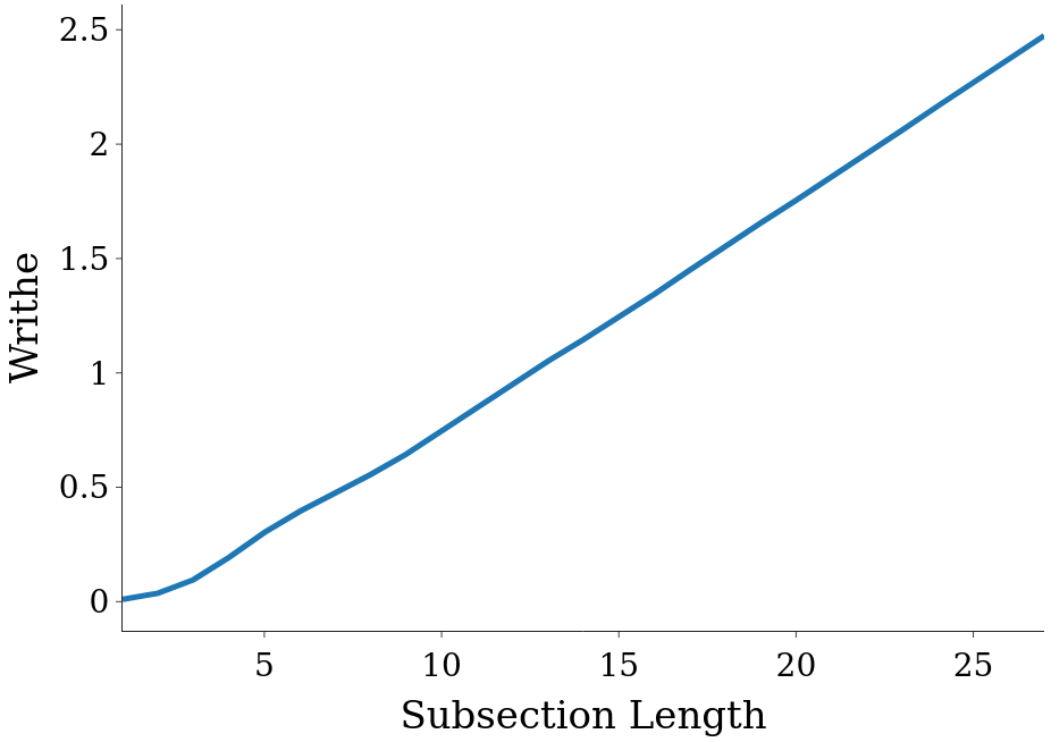


Figure 2.13: The writhe profile of the helical curve in Figure 2.12

proteins that share a specific writhe profile. For example, the writhe profile for a helical curve will grow linearly, due to the consistent handedness of coiling. Indeed, the writhe per turn for a helix of pitch P and radius R (as seen in fig. 2.12) is given by [66]

$$Wr_{\text{turn}} = 1 - \frac{P}{\sqrt{P^2 + 4\pi^2 R^2}}. \tag{2.16}$$

This is clearly seen in the writhe profile in Figure 2.13 which grows linearly to a value of 2.47. For this helical curve with radius of 0.6 and pitch of $\frac{1}{2\pi}$ the writhe per turn is 0.615, which gives a writhe of 2.46 for the four complete turns seen.

2.3.3 The length scaling complexity of the writhe

The linear nature of the writhe profile for a helix is due to its highly ordered state, with each subsequent coil adding writhe uniformly. For this reason, one may ask whether this represents an ideal conformation in terms of a maximal build up of writhe. We will look to answer this question in thesis, and in doing so, build on the existing literature on the limits on entanglement as measured by the writhe.

The earliest work in this area [67] found that the absolute value of the writhe of a self-avoiding walk scales like n^{α} where n is the length of the walk and $\alpha \approx 0.5$. This technique was applied to a set of 197 proteins in [68], determining an empirical power law of $0.045n^{1.4}$ for the scaling of the acn of a protein with n amino acid residues. To develop this further, Arteca compared the scaling relationship for proteins of largely different lengths in [59]. The main conclusion here is that for larger proteins (> 300 residues), the scaling exponent for the average number of crossings is smaller than for proteins with shorter chain lengths. Arteca proposes that this is due to the difference in secondary structure content between these two regimes. In particular, β strands are almost twice as prevalent in proteins with fewer than 300 residues than in those with longer chains. This allows the shorter chains to form more compact structures, with longer chains containing larger super-secondary structure. Arteca also considered the scaling of entanglement for sub chains within a set of 904 proteins in [69], and found empirically that the acn of a sub chain of length n scales similarly to that of a protein whose full chain length is n. The relationships

between global entanglement complexity and secondary structure will be key to our work in chapter 3, where we apply these concepts to a significantly larger dataset.

In [70] a theoretical upper bound on the writhe of smooth thick knots is found, given by:

$$|Wr(K)| \leq \frac{1}{4}\left(\frac{L}{R}\right)^{4/3}, \tag{2.17}$$

where L is the knot K's length and R its radial thickness. In this work, an upper bound on the helicity of unit vector fields is first derived, and then converted into a bound on the writhe of a knot using the so-called "bridge" theorem of [71] which relates these two quantities. Since we are interested in the discrete backbone curve, this bound on smooth knots does not immediately apply to proteins. Similarly, despite attempts such as [72], the radius of the backbone curve is not well defined. Despite these two difficulties, we will see that this theoretical bound provides a good benchmark for studying protein entanglement.

The length scaling of *acn* for equilateral random walks was shown to grow like $\mathcal{O}(n \log n)$ in [73]. In subsequent work [74], this scaling was seen to describe the length-dependent complexity of *acn* for larger protein backbones within a sample of 2230. However a different power law was required to describe the scaling for the small backbones in their sample. When an equilateral random walk is constrained to a convex confined space, the *acn* growth is on the order of $\mathcal{O}(n^2)$ and the writhe has growth on the order of $\mathcal{O}(\sqrt{n})$ [75], though no attempt was made to connect this to the scaling in entanglement complexity for protein backbone curves. The second Vassiliev measure, another measure of entanglement derived from a form of the Gauss self-linking integral, was shown to have order $\mathcal{O}(n)$ scaling for uniform random walks in [76], falling between the writhe and *acn* growth seen in earlier work. Due to the strict local geometrical constraints on the protein backbone, we may expect its growth in complexity to be limited compared to random walks. Indeed, we will see in chapter 3 the relationship between the *acn* of protein backbones and their secondary structure, and the super-secondary structure associated with maximal *acn* conformations. In particular, we find a complexity scaling law which which is consistent across all length scales.

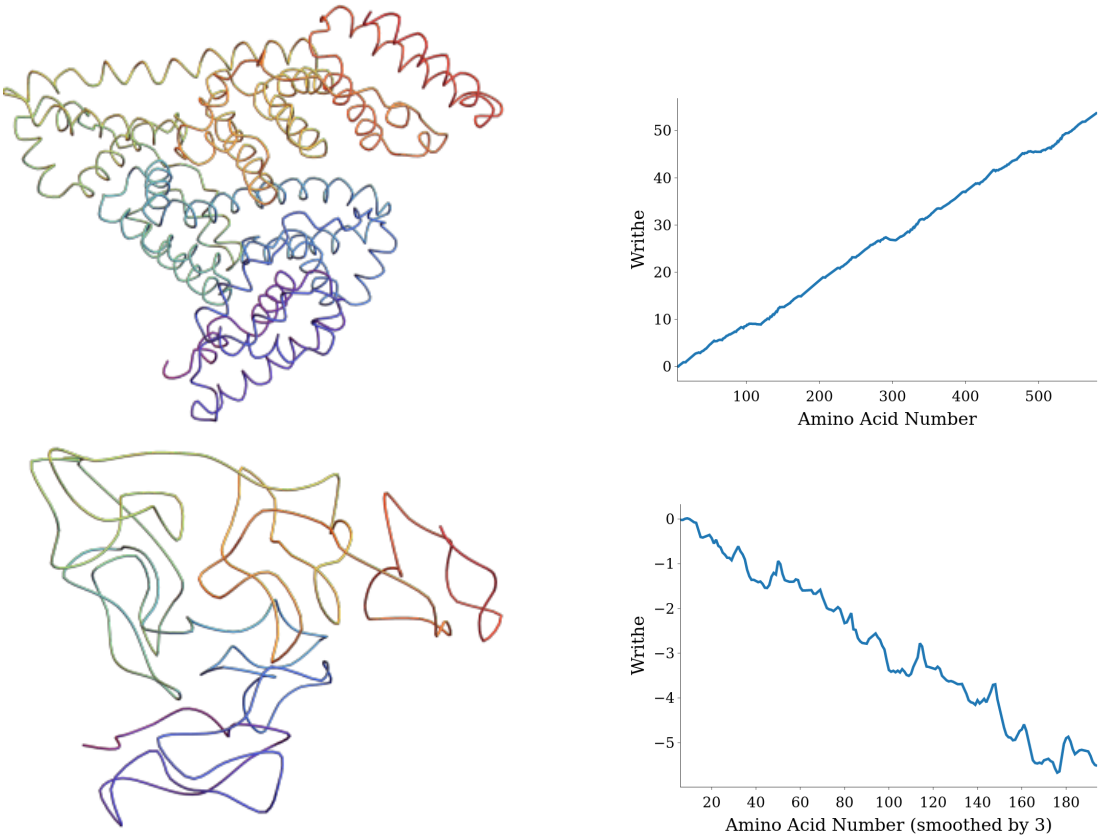


Figure 2.14: Illustrations of the effect of smoothing the backbone to avoid secondary structure writhe. Panel (a), the Cα backbone of Bovine Serum Albumin (PDB entry: 3V03). Panel (b) the writhe as a function of length for the backbone curve in (a). Panel (c) is the backbone curve in (a), sampled every three amino acids. Panel (d) is the writhe as a function of length for the curve in (c).

2.4 Smoothing the backbone curve

2.4.1 A motivating example

When topological measures are applied to the study of the tertiary structure, many researchers have found it useful to smooth the backbone curve in some way. For example, the locally helical nature of secondary structures contribute large amounts of writhe, while not impacting the global fold. Indeed, in Figure 2.14 we see an example of the quite drastic effect backbone smoothing can have on topological measures. In Figure 2.14A we see a cartoon representation of the Cα backbone curve of Bovine Serum Albumin (PDB entry: 3V03), a protein which is dominantly α-helical. In Figure 2.14B we see the writhe profile for this backbone curve, which has positive linear growth as expected for the dominantly helical conformation. In

Figure 2.14C we see the curve produced by sampling every third amino acid of the backbone curve in Figure 2.14A. The writhe profile of this curve, shown in Figure 2.14D, is remarkably different. The negative linear growth here is indicative of a large scale negatively coiled helical structure. By removing the contribution of the positively coiled α-helices, we uncover the underlying negatively coiled helical arrangement of these secondary structure elements. Though this is a quite crude approach, we can already see the powerful impacts of smoothing the backbone.

2.4.2 Existing methods

We now discuss some of the more sophisticated approaches taken to backbone smoothing and their uses. Hinds proposes a lattice model whose vertices can represent many residues in [77]. Despite losing the secondary structure, this method is able to predict possible conformations of five structurally dissimilar small proteins. In [78], the backbone is smoothed by averaging coordinates along a sliding 7-residue window as a method of visually highlighting the similarities detected by the structural measure used in this work. A more extreme smoothing method is introduced in [79], where points are removed iteratively to reduce a polymer to a minimal representation of its knot type. This massively speeds up the computation of the knot type; for instance, an unknotted protein will be reduced to simply the edge connecting its endpoints. However, the obvious disadvantage of this approach is the loss of any recognisable protein-like structure. This curve reduction technique was modified in [80], where instead or removing points, they are moved in order to minimise the total curvature of the curve. This allows for the removal of the helical nature of secondary structure whilst maintaining a recognisable tertiary fold. This curvature minimising based smoothing is applied in [81] before computing the Gauss integral measures from [61].

All of the above smoothing techniques can broadly be put into two categories, those that smooth by removing/reducing the number of points, and those that move points in order to change some geometrical property of the curve. The former is usually favoured when a reduction in complexity is needed for computing some structural measure, for example, the computationally intensive knot polynomial

calculations. However, by removing points you may lose recognisable tertiary structure, making interpretation of structural information difficult. Smoothing methods that simply move points instead of removing them do not suffer from this. They do clearly however have the inverse problem; with no reduction in the number of points computations are no less costly. In Chapter 3 we will introduce a smoothing technique which falls between these two categories, reducing the number of points whilst maintaining a recognisable tertiary fold.

2.5 Discussion

In this chapter, we have provided the mathematical background to study the structure of proteins on various scales. We surveyed the current literature and gave justification for our choices of geometrical and topological tools to describe protein backbone curves.

We first discussed the heavily constrained geometry of the backbone curve, especially with α-helices and β-strands. These constraints are due to the Ramachandran angles for all atomistic structures. Since our discrete backbone curve model comprises solely the central Cα atoms, the Ramachandran angles cannot be directly computed. To address this, we defined discrete curvature and torsion, which are analogous to the Ramachandran angles in this context. These quantities are tightly constrained within secondary structures, justifying their use as parameters in a coarse-grained backbone model. Compared to other approaches taken to modelling the backbone curve, the constrained backbone algorithm is unique in its minimal parameter approach to modelling the backbone for fitting to BioSAXS data.

We highlighted one limitation of the constrained backbone algorithm as presented in [23]. Namely, though the curvature-torsion parameterisation ensures a geometrically realistic local structure, it has little control on the biological plausibility of the tertiary structure. To address this problem, we require a method for comparing and classifying tertiary structures. We discussed some of the existing work in this area, which broadly falls into two categories; alignment based methods, which we showed to have some drawbacks critical to our specific problem, and topological methods,

which addresses these drawbacks. Knot theoretic approaches to protein structural comparison form the majority of the topological work in this area. Though these tools are incredibly powerful, we introduced the writhe and justify our use of it in building our similarity measures. Namely, we highlighted its fundamental link to helical geometries which will become useful in later chapters.

Inspired by previous work in this area, we defined a writhe fingerprint and from this a writhe profile. We introduced these as tools to study the change in writhe for specific subsections of a curve, which will be useful for identifying super secondary structural motifs in proteins. The writhe fingerprints and profiles of some simple exemplar curves are shown, with these specific motifs introduced as they will appear in the following chapter where we show that they are common to many protein backbones. We finally discussed the notion of backbone smoothing, and show through an example its importance when computing topological measures of similarity. We discussed some of the existing work in this area, highlighting where each approach does not meet our specific needs.

CHAPTER 3

The SKMT Algorithm for comparing underlying protein entanglement

In this chapter we develop a novel method for smoothing a protein backbone curve into a minimal representation of its global entanglement. The motivation for smoothing the backbone in this way is to highlight the rigidity of secondary structural elements when considered as part of the global fold. This in turn informs the specific way in which to define the length of this smoothed backbone representation.

We then study the distribution of writhe for these smoothed backbones, providing clear bounds on the expected amount of backbone entanglement relative to the amount of secondary structure. By studying proteins whose writhe approaches these bounds, we uncover large scale helical geometries present across a diverse range of proteins. These helical geometries exists as global conformations as well as super secondary motifs, with a consistent scale.

We investigate the strength of these bounds by considering a common pitfall of experimentally determined structures, namely the potential for Cα atoms to be missing from the N-terminus. We find that this has little effect on the distribution of writhe relative to secondary structure. We also further justify our choice of length by considering the distribution of writhe against other possible choices. We show

that these alternative choices are unable to clearly capture a consistent growth in entanglement complexity across all length scales.

We conclude the chapter with two brief tangential curiosities. First, by plotting the distribution of writhe for a sample of trefoil knotted proteins, we show that there is little correlation between the knottedness of a protein backbone and the writhe of its smoothed representation. This highlights the unique information captured when considering the writhe of these smoothed backbone curves. Secondly, we introduce a type of curve whose arrangement of consecutive helical loops minimises self entanglement. We show that this conformation has an easy to detect writhe profile motif, which we find is present amongst many protein backbones.

The results presented in this chapter formed the basis for the following publication [3].

3.1 The SKMT algorithm

3.1.1 Motivation

The locally helical nature of secondary structure elements (SSEs) means that they contribute large amounts of writhe despite having little effect on global entanglement. Due to their tightly constrained local geometry, as discussed in Chapter 2, the contribution to the tertiary fold of α helices and β strands is essentially that of a straight edge. As a result, before applying a structural similarity measure built on the writhe, it is important to smooth the backbone curve in some manner to remove the locally helical structure. This notion has been widely addressed in the literature, for example, by uniform averaging of residues as in [63] or by minimising total curvature as in [80]. In what follows, however, we show these approaches miss the close relationship between secondary structure content and possible global entanglement.

Given the rigid nature of secondary structure elements on the global scale, the naive smoothing approach would be to replace each secondary structure with a single edge connecting their start and end points, such that the length of the smoothed representative curve is twice the number of SSEs. In the majority of cases for α

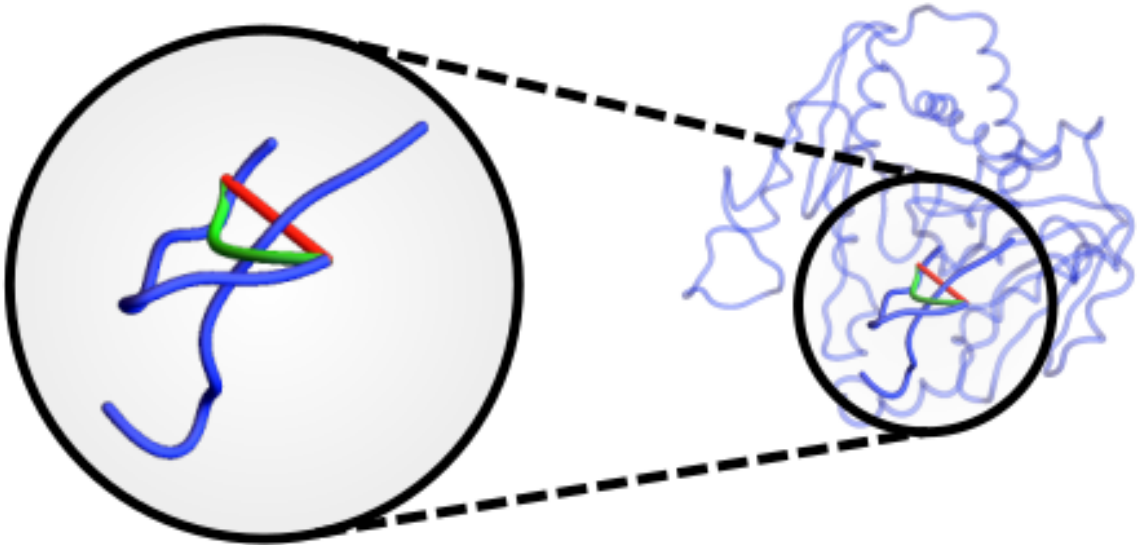


Figure 3.1: In blue, the backbone curve of the trefoil knotted acetylornithine transcarbamylase (PDB 3KZK). Highlighted is the threading of the C-terminal strand through a linker subsection that is essential for the trefoil knotting. In red, we see the straight edge between the start and end points for this linker subsection, which passes beneath the C-terminal strand. In green, we see a minimal sub-curve which maintains the essential entanglement.

helices and β strands, this is in fact a fairly safe approach. However, since the linker subsections are much more structurally variable, this approach can cause some issues. In Figure 3.1, we highlight a linker subsection that plays an important role in the trefoil knotting of the acetylornithine transcarbamylase protein. The linker loops back on itself, and is threaded by the C-terminal strand in order to complete the knotting. If we were to replace it by the single straight edge connecting its endpoints as for the other SSEs, we would fundamentally change the topology of this curve (shown in red). Instead, we should reduce this linker to a minimal curve that maintains this essential entanglement (similar to that shown in green). To achieve this, we can adapt the KMT algorithm [44, 79] to act solely on this linker subsection. In the standard KMT algorithm, if the triangle defined by three consecutive points of a curve is not intersected by any other edge of the curve, the middle point of these three is removed. This triangle reduction is then repeated to produce the minimal representation of the curve's topology. For our approach, we will restrict the triangle reduction to act within secondary structure elements, thereby preserving a recognisable tertiary structure. For this reason, we will call this

the SKMT (Secondary KMT) algorithm, and refer to its output as SKMT smoothed backbone curves.

3.1.2 Algorithm

We take as input the coordinates of the $C\alpha$ backbone curve from the PDB file and a secondary structure assignment, *eg* from PSIPRED [82]. We construct the SKMT smoothed backbone curve as follows. One fundamental aspect of the algorithm is determining if the triangle defined by three points is intersected by the edge defined by some other two points. This is in fact a very common problem in the field of computer graphics, and for our solution we turn to the algorithm described in [83]. For this, we define the signed volume S_V of the tetrahedron defined by four points $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$ by

$$S_V(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4) = \frac{((\mathbf{x}_2 - \mathbf{x}_1) \times (\mathbf{x}_3 - \mathbf{x}_1)) \cdot (\mathbf{x}_4 - \mathbf{x}_1)}{6} \tag{3.1}$$

Let $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$ be the vertices of a triangle, and $\mathbf{q}_1, \mathbf{q}_2$ be two points in $\mathbb{R}^3$. Then the edge defined by $\mathbf{q}_1$ and $\mathbf{q}_2$ intersects the triangle $\mathbf{p}_1\mathbf{p}_2\mathbf{p}_3$ if both of the following are true:

- $S_V(\mathbf{q}_1, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3)$ and $S_V(\mathbf{q}_2, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3)$ have different sign
- $S_V(\mathbf{q}_1, \mathbf{q}_2, \mathbf{p}_1, \mathbf{p}_2)$, $S_V(\mathbf{q}_1, \mathbf{q}_2, \mathbf{p}_2, \mathbf{p}_3)$, and $S_V(\mathbf{q}_1, \mathbf{q}_2, \mathbf{p}_3, \mathbf{p}_1)$ all have the same sign.

As can be seen in Figure 3.2, this intersection check is performed for every triangle within each SSE, against all other edges of the backbone curve. We are therefore preserving any non-local entanglement of secondary structures such as knotting or slip-knotting. This preservation of super secondary structure is what differentiates this from applications of the original KMT algorithm of [44,79]. An example protein backbone curve is shown alongside its SKMT smoothed representation in Figure 3.3.

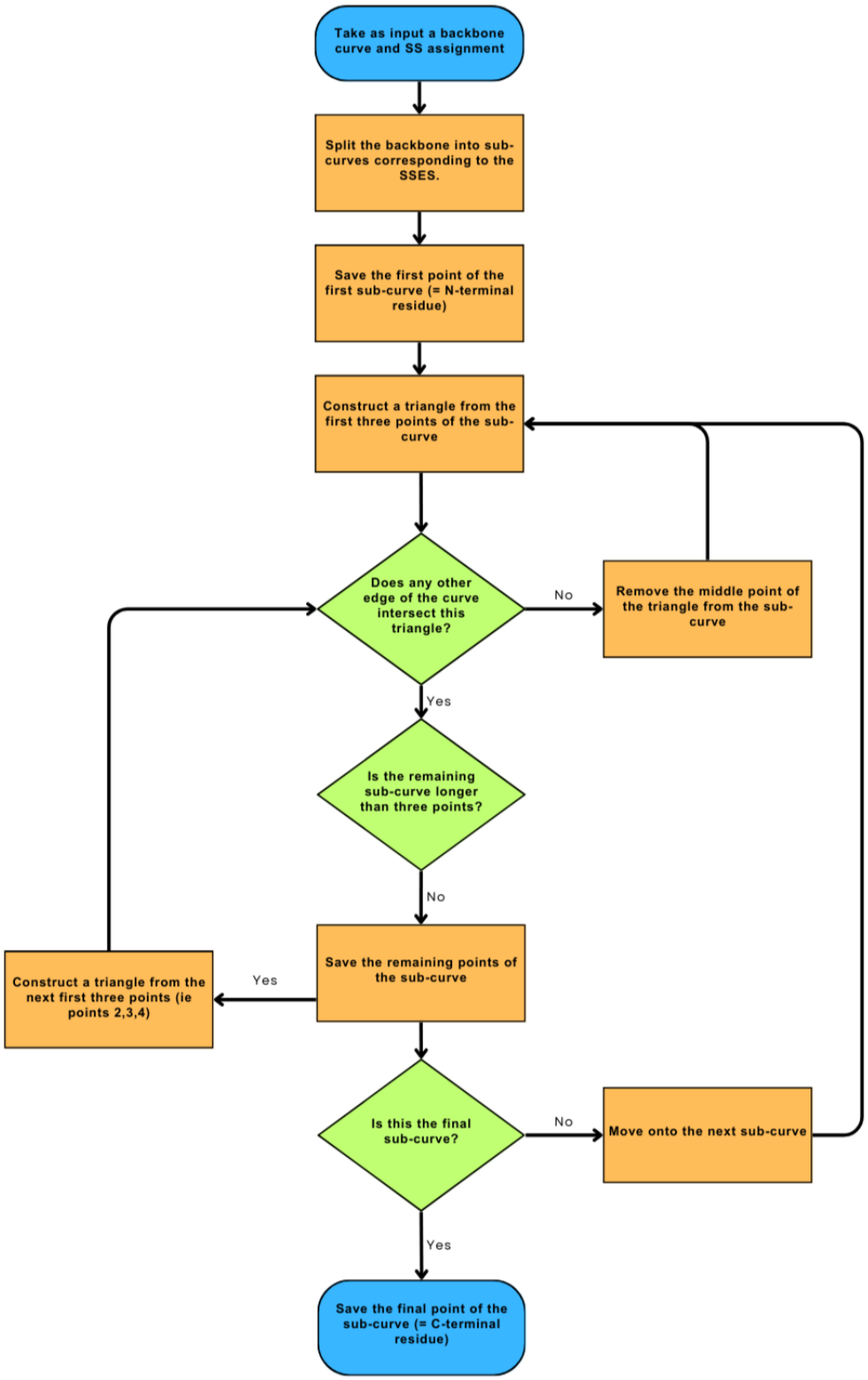


Figure 3.2: A flow chart describing the SKMT algorithm. Note that any SSE sub-curve that is fewer than three points long will be automatically saved.

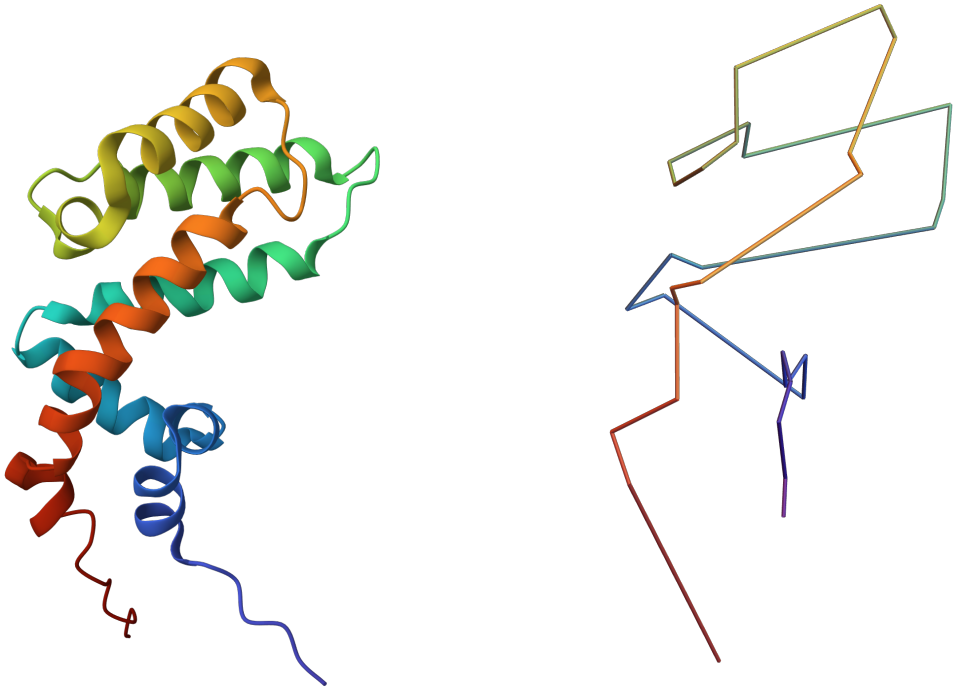
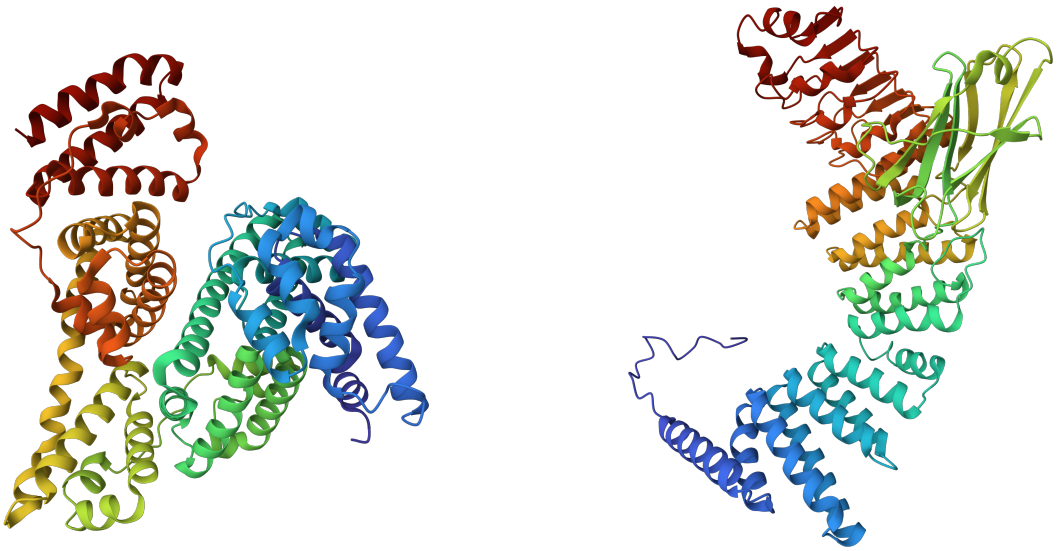


Figure 3.3: The cartoon backbone of the RGS-homologous domain of Axin alongside its SKMT smoothed representation.

3.1.3 SKMT length

In what follows, we define the length of a protein as the number of points of its SKMT curve. For relatively unentangled backbones, this roughly coincides with the number of distinct SSEs, since each SSE will be replaced by a single edge in the SKMT algorithm. However, for backbones with more complex non-local entanglement (such as the case in Figure 3.1), this will be greater than the number of SSEs due to the inclusion of the points necessary to preserve the topology. This choice allows us to give clear bounds on the growth in complexity of the entanglement of the backbone with respect to its secondary structure content. We will show later that other choices of length do not give clear and consistent bounds on the length dependent scaling of entanglement.

To further illustrate the motivation for this choice, consider the two proteins shown in Figure 3.4. The structure shown in Figure 3.4A is Bovine Serum Albumin (BSA), which consists of 583 amino acids with 78 relatively short secondary structure elements. The structure shown in Figure 3.4B is Rab Geranylgeranyltrans-



A The cartoon representation of the backbone of Bovine Serum Albumin (PDB entry: 3V03).

B The cartoon representation of the backbone of Rab Geranylgeranyltransferase (PDB entry: 1DCE).

Figure 3.4: Two proteins with similar primary sequence length, but significantly different number of secondary structure sections and therefore possible complexity.

ferase (RabG), which has 567 amino acids but just 59 secondary structure elements. Despite having roughly the same amount of amino acids, the *acn* of the SKMT smoothed backbone of BSA is 41.9, roughly 2 times greater than that of RabG at 20.1. This coincides with BSA having 1.3 times more secondary structure content than RabG. Indeed, the fewer and therefore longer inflexible sections in RabG act as a barrier to entanglement. This example highlights that the global entanglement is more closely linked to secondary structure than the number of amino acids, making the SKMT length a natural choice.

3.2 Length constraints on the writhe of proteins

Now that we have an effective method of smoothing the backbone curves to uncover their global self entanglement, we will study the range of possible values this entanglement can take. Using the same representative sample of proteins as defined in Chapter 2, we compute the writhe of the SKMT smoothed representation of each

protein's backbone curve, and plot it against its SKMT length. The distribution of writhe against length is shown for this sample of proteins in Figure 3.5, along with four exemplar protein backbones.

We highlight some of the critical features of this plot for further discussion

1. 99.9% of all values lie within the curves

$$\pm\frac{1}{4}\left(\frac{L}{R}\right)^{4/3}, \tag{3.2}$$

   the theoretical knot bound from [70], with a radius $R = 2.7$ (coinciding with the mean "tube" thickness of 2.7Å found in [72] for minimal Cα triplet radii).

2. For larger proteins, $L > 35$, the Wr values increasingly fail to get close to this limit. We find 97.9% of the structures fit within a linear bound $0.12L$.

3. There exists proteins with low or even close to zero net writhe consistent across all scales.

On the first point, as discussed the theoretical knot bound of [70] was developed for smooth thick knots. Plainly, our SKMT backbones are not smooth, and have no natural definition of a radius. As a result, this bound does not strictly apply, and can be seen to be a huge overestimation for larger proteins. This result is consistent with the findings of [68], further strengthening the argument that the entanglement of the backbone curve is dependent on the secondary structure. Larger proteins have the potential for larger secondary structure sections, which act as a barrier to more complex entanglement. This is not the case for the smooth knots to which the theoretical bound applies, that have no restrictions on their local geometry. All outliers to the theoretical knot bound have $L \leq 17$. For example, the backbone shown in Figure 3.5A has its five secondary structure elements coiled with consistent chirality, leading to a sharp build-up of writhe. Maintaining this systematic entanglement for proteins much larger than this with more secondary structure sections is difficult.

For a tighter bound on the possible writhe values of larger proteins, we find that a linear curve with gradient 0.12 is consistently tight to the data across all lengths.

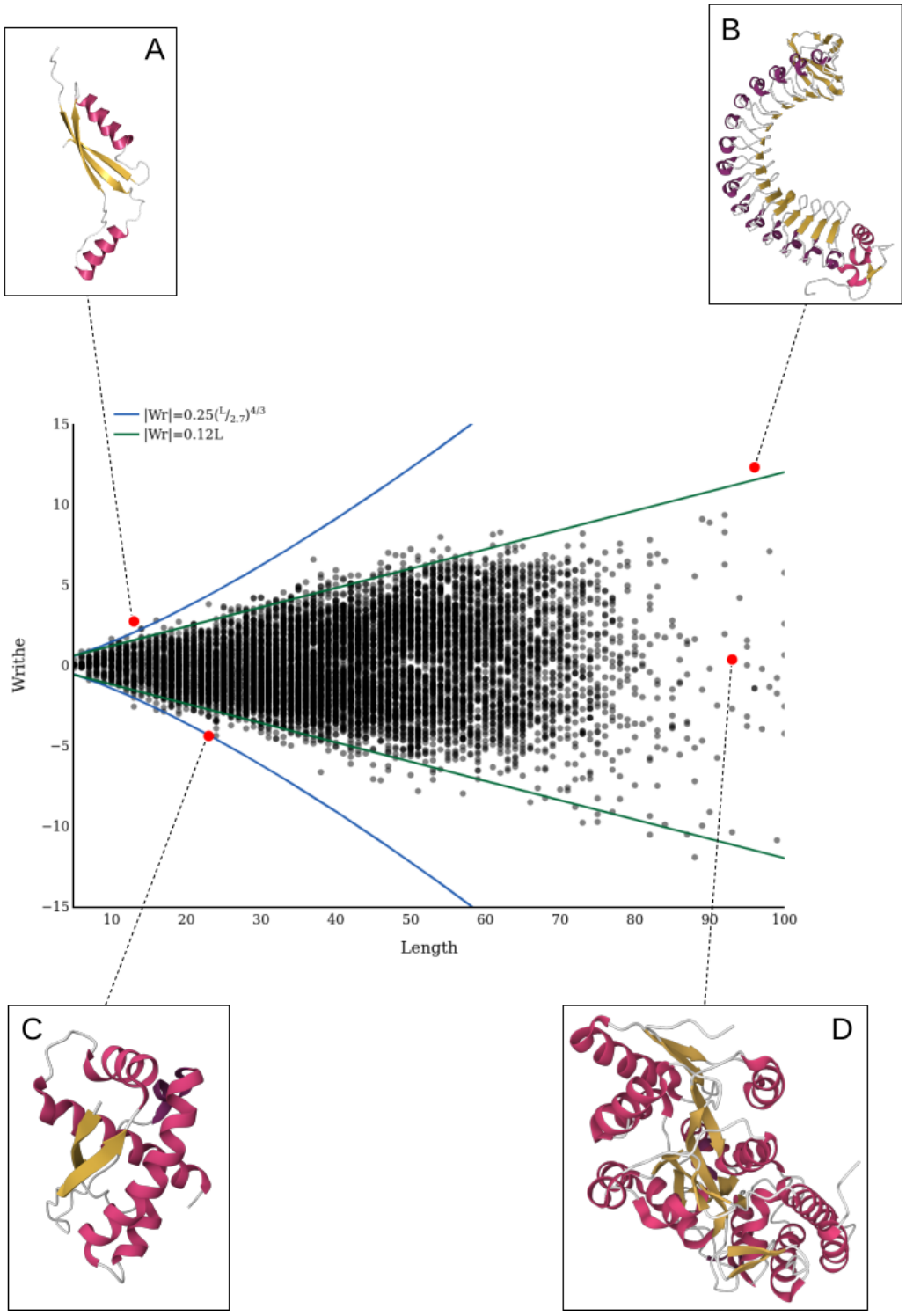
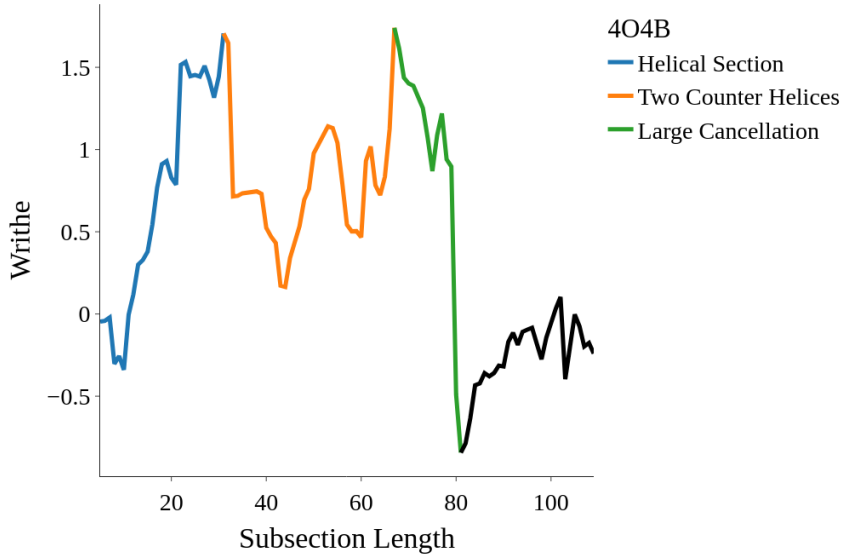


Figure 3.5: The distribution of writhe for a representative sample of > 10000 proteins from the PDB in black. In blue we see the theoretical writhe bound [70] with $R = 2.7$. In green, we see linear growth in writhe with a gradient of 0.12. Inset: A: PDB 1VQ3. B: PDB 2OMZ. C: PDB 2RH3. D: PDB 4O4B.

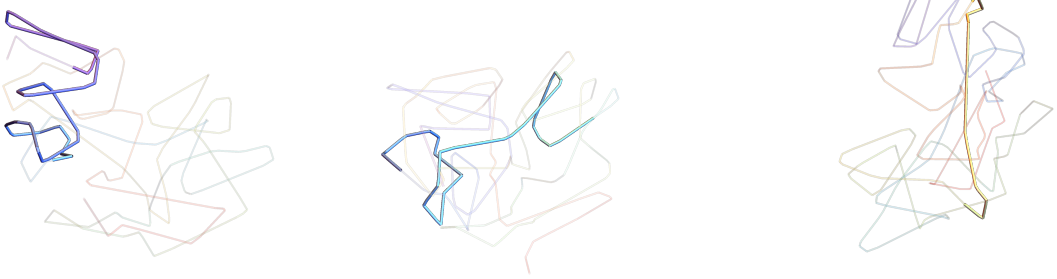
As we have seen, helices have linear growth in writhe with respect to their length, suggesting that large-scale helical geometries act as maximal entanglement states for protein backbones. We will investigate this further in Section 3.2.2. An example of the highly organised helical structure required for maximal writhe is highlighted in Figure 3.5B. This protein has an alternating $\alpha - \beta$ secondary structure motif, organized into a helix on a horseshoe-shaped central axis.

The protein highlighted in Figure 3.5 C is a relatively small trefoil knotted protein. Due to its knottedness, this protein exhibits super-linear growth in writhe relative to its secondary structure. In fact, its writhe lies almost perfectly on the theoretical knot bound. We will see later in Section 3.4.1 that in general the writhe of the SKMT curve does not correlate with the knottedness of the protein's backbone. In this case however, this conformation is an example of a shallow-knotted protein [84]. That is, a protein where removing just a few residues from the ends would undo the knotting. Since the knot requires almost the entirety of the backbone, we can see why this conformation is maximally entangled with respect to its length.

On the final point, zero writhe structures are present across all length scales. These however are not trivially entangled backbones. As an example, in Figure 3.6 we plot the writhe profile of the backbone highlighted in Figure 3.5D. In Figure 3.6A we see that $Wr(\mathcal{C}_{1n})$ peaks just above 1.5, which corresponds to the helically coiled substructure $\mathcal{C}_{1,29}$ highlighted in Figure 3.6B. The appearance of large scale helices as maximally entangled subsections will be discussed further in Section 3.2.2. In Figure 3.6C we see two counter-helical loops that do not contribute net writhe due to cancellation in the signed sum. This specific writhe profile prompted further study, which is discussed in Section 3.4.2. Finally, the long subsection passing through the rest of the structure highlighted in Figure 3.6D leads to a cancellation of the previously accumulated writhe. Given these structures identified within this writhe profile are not unique to this protein, we develop routines to identify if they are present within a protein of interest.



A The writhe profile of the SKMT smoothed backbone of an Inositol hexakisphosphate kinase. (PDB entry 4O4B)



B The SKMT smoothed backbone of 4O4B, with the helical subdomain highlighted.

C The SKMT smoothed backbone of 4O4B, with the two counter helical domains highlighted.

D The SKMT smoothed backbone of 4O4B, with the subdomain causing large cancellation highlighted.

Figure 3.6: An example of the potential domains present in a complex entangled yet net zero writhe structure.

3.2.1 Identifying helical super-secondary structure.

In order to identify helical substructures, we need to identify regions of the writhe profile that could be well approximated by a linear curve with non-zero gradient. To do this, we use the following routine

1. We first perform a LOWESS (locally weighted scatterplot smoothing) on the writhe data.

2. For all subsections of length greater than 10, we compute the gradient of the writhe profile for this subsection.

3. If this gradient is larger than 0.06 (*ie* within 50% of the maximally observed linear growth), this subsection is potentially helical.

4. For a potentially helical subsection $\mathcal{C}_{(i,j)}$, we check for a change in sign of the gradients $Wr(\mathcal{C}_{(i,i+k+1)}) - Wr(\mathcal{C}_{(i,i+k)})$ for $k = 1, \ldots, j - i - 1$. This ensures that we find subsections with consistent linear growth.

5. Finally, we output the largest disjoint subsections that satisfy the above criteria.

3.2.2 Investigating helical super-secondary structure.

As noted in [85], a key area of interest for researchers is the arrangement of many elements of secondary structure into commonly occurring motifs, referred to as super secondary structure. For example, in the protein highlighted in Figure 3.5B, the large scale helical geometry is formed from alternating α-helices and β-strands. This alternating $\alpha-\beta$ is a fundamental characteristic of CATH domains such as Rossmann folds [86] and TIM barrels [87]. In Figure 3.7 we plot the writhe profiles for both a Rossmann fold and TIM Barrel domain. One can see that the overall gradient is similar for both structures. Between sections 5 and 42 the writhe in both cases grows by about 3.2, which gives a gradient of 0.0865 (to 3.*s.f*). Each of these examples is known as a tandem repeat [88], which are of keen interest to researchers due to their protein-protein interactions and specific binding surfaces. By identifying large scale

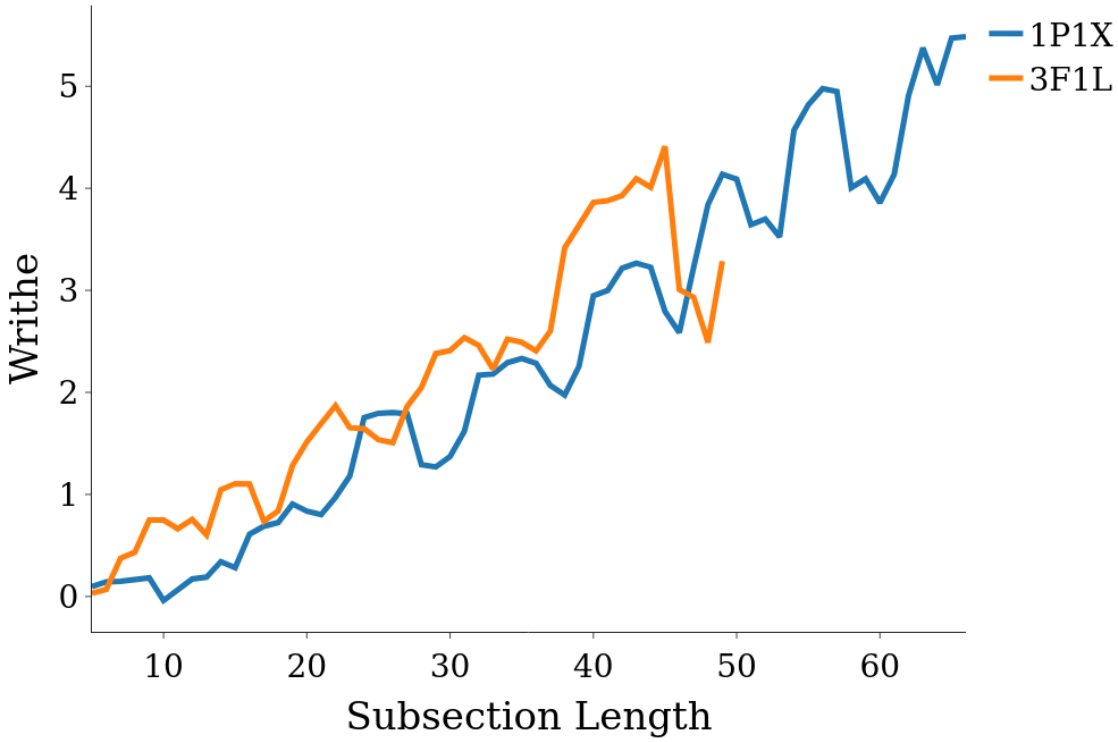


Figure 3.7: A comparison of the writhe profiles of two similarly helical protein structures. In blue, the TIM-Barrel domain 1P1X, in orange, the Rossmann fold domain 3F1L.

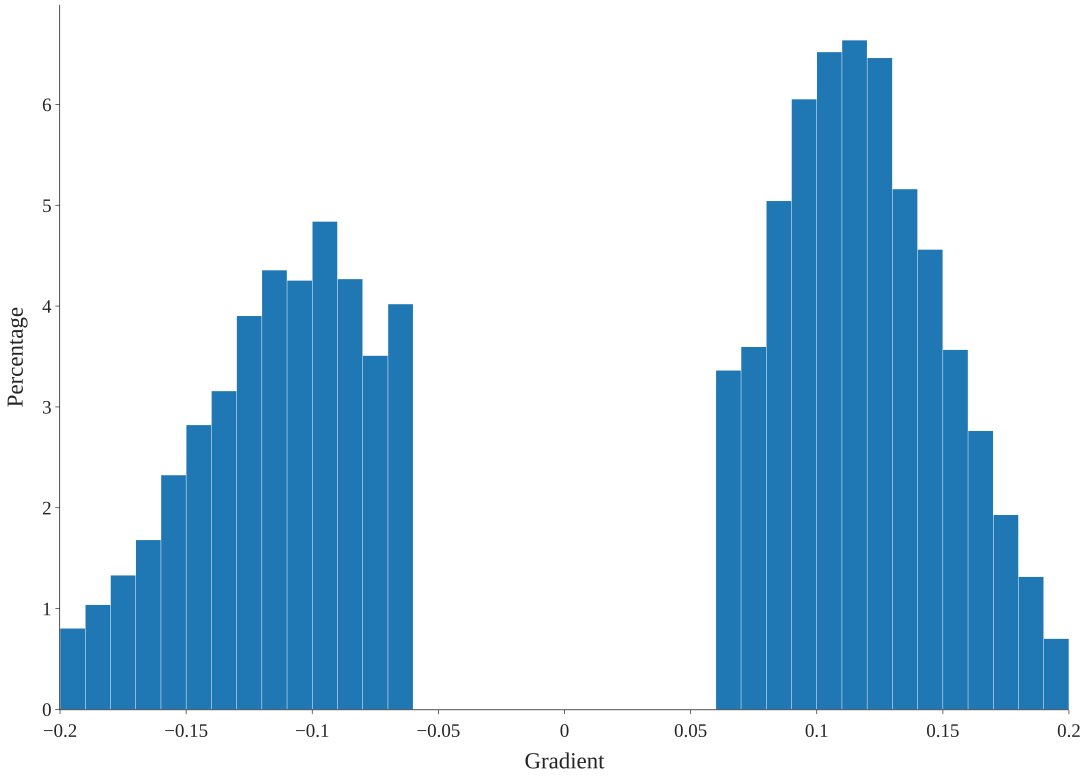


Figure 3.8: The distribution of gradients of the writhe profiles for helical subsections of the SKMT smoothed backbone curves.

linear growth in writhe, we can quickly identify helical geometries and therefore structures such as the examples presented above. To investigate the prevalence of helical super-secondary structure, we run the routine described in Section 3.2.1 across our dataset.

In Figure 3.8 we see the distribution of gradients of the writhe profiles for any identified helical subsections from the dataset. This distribution is bimodal, with peaks for values with magnitude between 0.09 to 0.13. There are more positively entangled helices and the helical subsections with gradients greater than 0.15 are relatively sparse. From this we can see that the helical geometries seen in Rossmann folds and TIM barrels, as well as those found in the exemplar proteins Figure 3.5B and D, are indicative of the dominant helical geometry found throughout our representative sample of the PDB. Looking at the distribution of lengths for these subsections in Figure 3.9, we see that over 60% of them have lengths less than 16. An SKMT length of 16 requires at least 8 secondary structure elements, suggesting

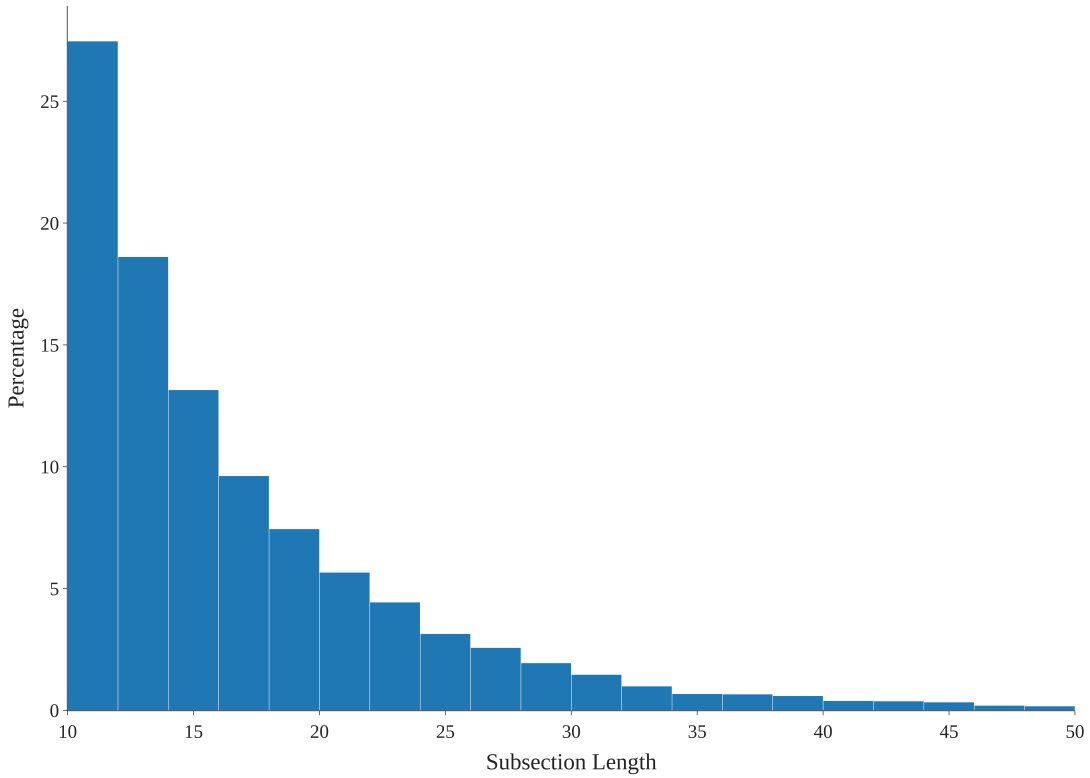


Figure 3.9: The distribution of the lengths of helical subsections of the SKMT smoothed backbone curves.

these helical subsections are on the same scale as super secondary structural motifs.

3.3 Tests that the bound is meaningful

3.3.1 Potential for missing N-terminal residues.

When working with experimental data, which comes with a degree of uncertainty, we should always be careful of the strength of our conclusions. For example, in the case of protein structural data, it is not uncommon for N-terminal residues to be missing from the PDB file. This may be due to the structural determination method used, for instance a purification tag may need to be attached to a protein chain during experiments. To account for this possibility in our results, we consider randomly cutting 10-20 residues from the start of the chain before applying the SKMT algorithm and then computing the writhe of the smoothed representation. In Figure 3.10 we plot the new distribution of writhe against length in red, overlaid with the original distribution of writhe against length in black. We see that there is little meaningful change to the distribution from cutting some residues from the start of the chain. Indeed, we now have 98.3% of the data falling within the linear bounding curve, compared to the 97.9% in the original case. This modification has little effect on the results mainly due to the strength of the method of backbone smoothing. Since 10-20 residues would roughly correspond to only one or two SSEs, we are cutting only one or two points from the start of our smoothed backbone curves. In doing so, we are doing little to change the nature of the overall entanglement of the curve. This point further reinforces the main argument of this work, that potential entanglement of the backbone is proportional to SSEs, and not the number of amino acids. This will be strengthened further by considering the distribution of writhe against these more classical definitions of length in the following subsections.

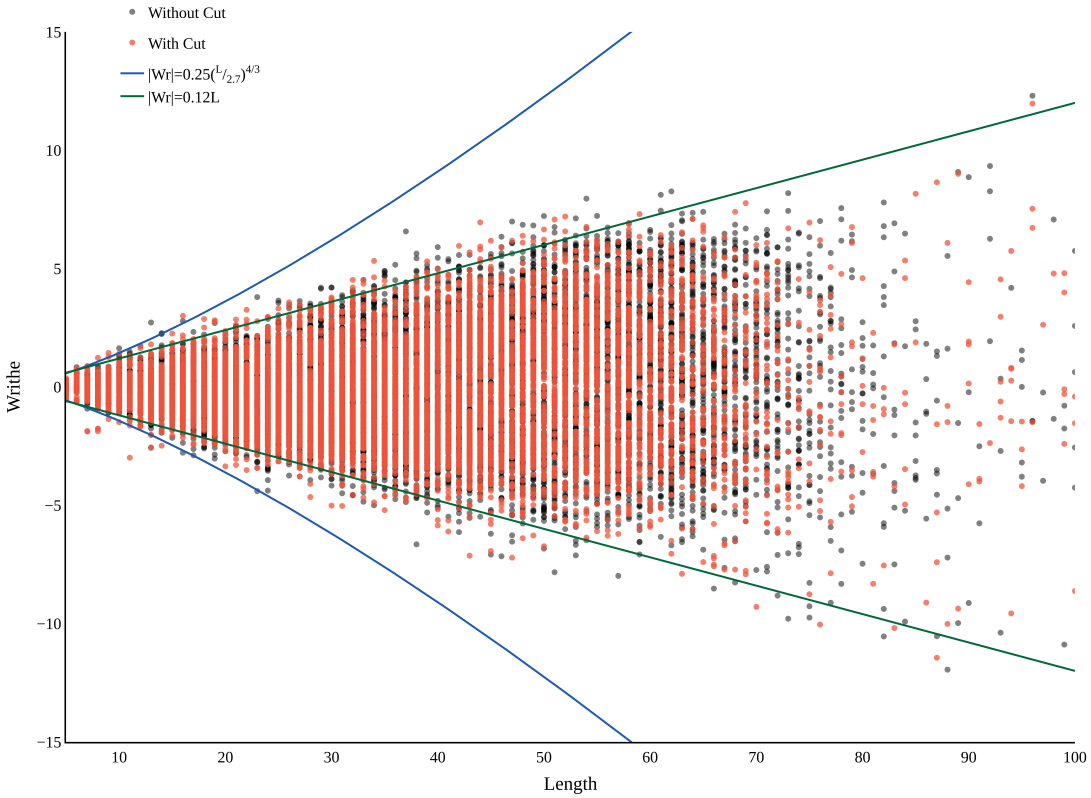


Figure 3.10: In red, the distribution of writhe against length for the SKMT backbone curves of proteins which have been modified to randomly cut between 10-20 of their N-terminal residues. This is overlaid with the distribution of writhe against length for the original SKMT smoothed backbone curves in black.

3.3.2 The best choice of length for entanglement scaling relationships.

To investigate further the impact of our choice of length on our results, we look at some of the other definitions that we could have taken. These include, but are not limited to

1. The arclength of the SKMT curve.
2. The number of amino acids of the protein.
3. The length of another smoothed curve (*eg* uniformly sampling every n amino acids).

To investigate this, we plot the writhe distribution for each of these definitions of length, and then fit a linear bounding curve to contain as close to the same portion of data (97.9%) as in the original case. With this, we see that we cannot find as clear and consistent a length-scaling relationship in each of these cases

Firstly, we consider the distribution against the arclength of the SKMT curve in Figure 3.11. Since this definition of length will be proportionate to our own, we expect the distribution to be broadly similar. Though this is the case, we see that the linear bound is less tight for lengths less than 300. As a heuristic measure of this tightness, we consider the average distance between the nearest point to the line across all lengths. We find that this is almost three times greater in the arclength case at 3.1 compared to 1.2 for our original distribution. By considering the arclength of the SKMT smoothed curve, the scaling of its entanglement is actually less clear than by simply considering the number of points of this curve.

Next, we consider the distribution of writhe against the number of amino acids of the respective protein in Figure 3.12. In this case, the poor fit of the proportional linear bounding curve is even more visually apparent, especially for small proteins. In terms of our heuristic "tightness of fit" measure, the average closest distance is 2.9 in this case. By only considering the number of amino acids and not their arrangement into secondary structure, we are losing some information in relation to the tertiary entanglement. For example, in Figure 3.13 we see the cartoon repre-

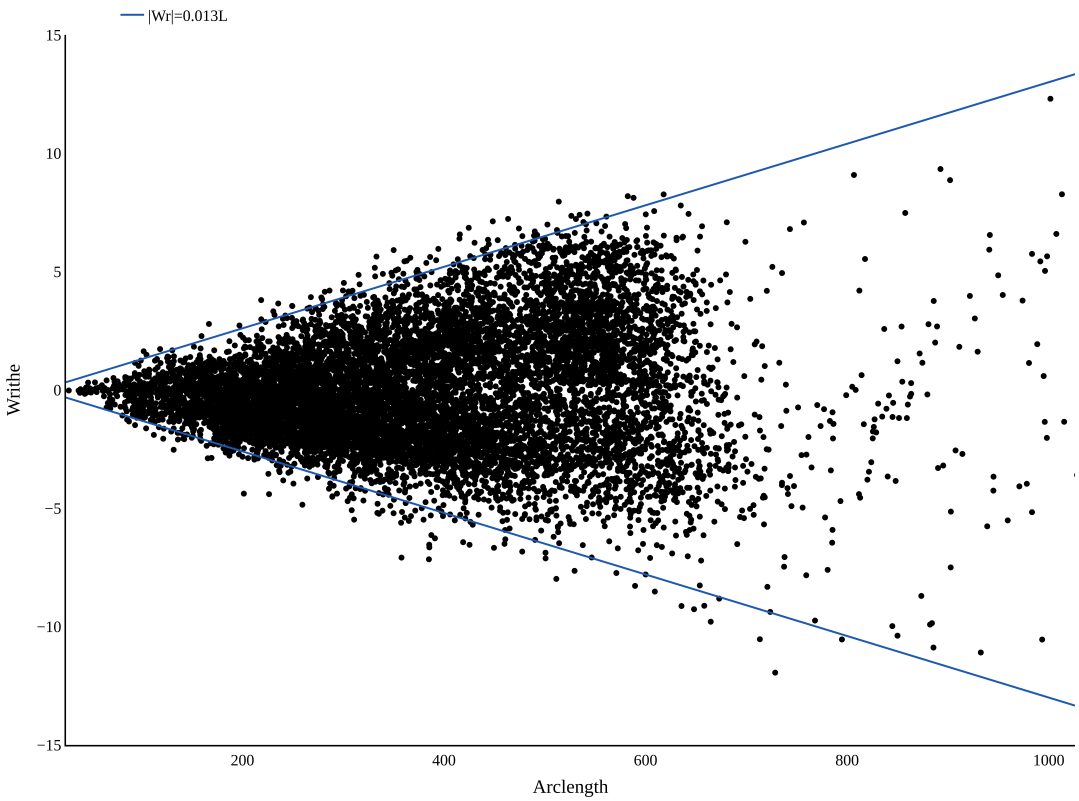


Figure 3.11: The distribution of writhe against the arclength of the SKMT smoothed backbone curves.

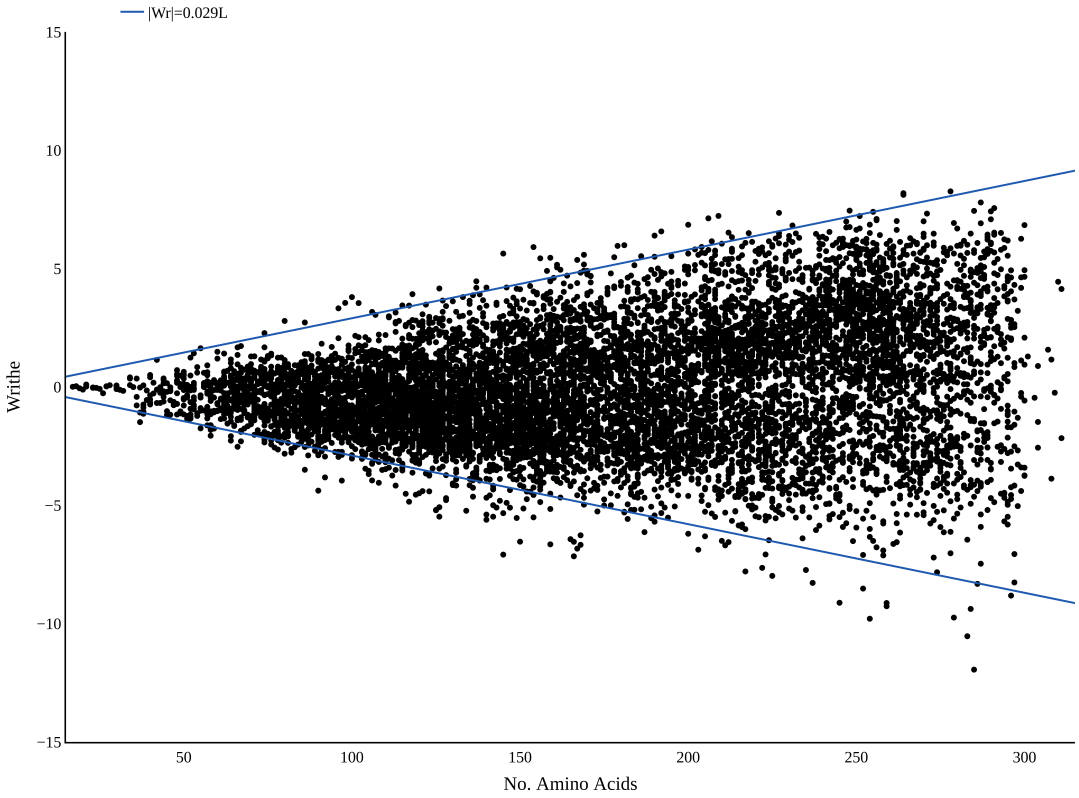


Figure 3.12: The distribution of writhe of the SKMT smoothed backbone curves against the number of amino acids of the respective protein.

A The cartoon representation of PDB entry 1VHF.

B The cartoon representation of PDB entry 4MT8.

Figure 3.13: An example of two proteins with similar number of amino acids but very different secondary structure content.

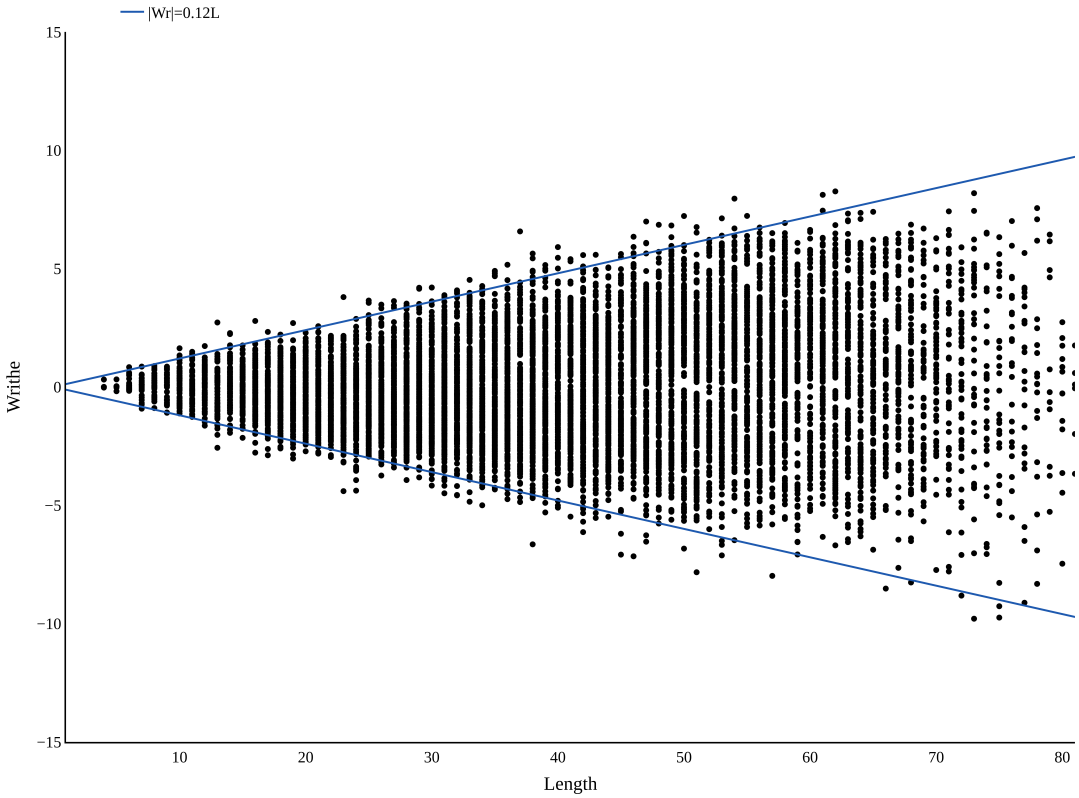


Figure 3.14: The distribution of writhe of the backbone curves, smoothed by uniformly sampling every 4th amino acid, against their length.

sentation of PDB entries 1VHF and 4MT8, consisting of 101 and 102 amino acids respectively. The SKMT smoothed backbone curve of 1VHF has writhe of -3.02 whereas the writhe of 4MT8 is 0.72. The visual difference between the two is clear: 1VHF has many relatively small SSEs forming a complex entanglement whereas 4MT8 is made up of just two long α-helices forming a hairpin topology. As a result, the magnitude of global entanglement is much greater for 1VHF. Put simply, it is very difficult for two long rigid rods to get entangled, as is the case for 4MT8. This example provides the clearest picture of our argument, that the range of possible entanglements is dictated by the size and number of the rigid secondary structure elements more than it is the number of amino acids.

We finally looked at the possibility of smoothing the backbone in a simpler way, by uniformly sampling every n^{th} amino acid. In Figure 3.14 we plot the writhe distribution against the number of points in the uniformly sampled curve for backbone curves sampled on every 4^{th} amino acid (similar results were observed for

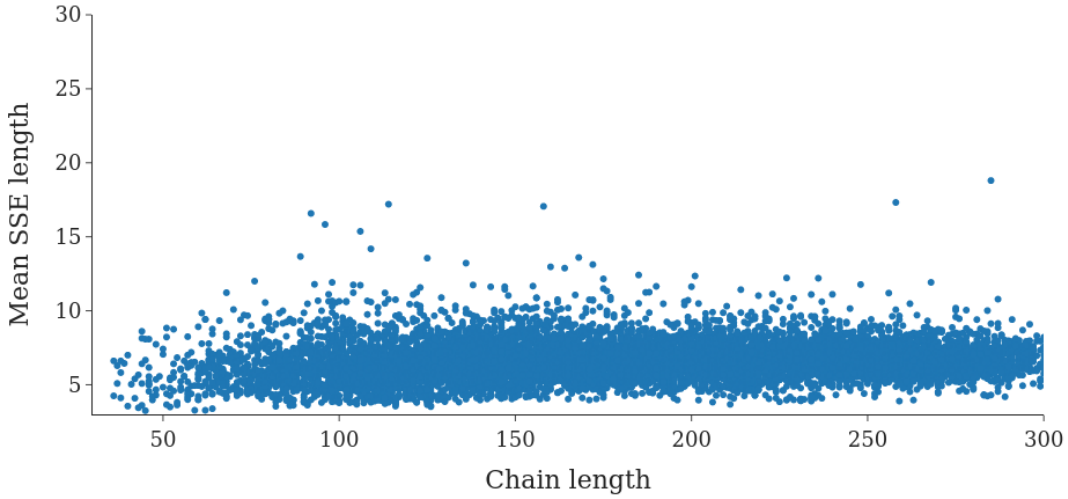


Figure 3.15: The average length of secondary structure elements against the length of the chain.

$n = 3, 5, 6$). A linear bounding curve with gradient 0.12 (as in the SKMT case) captures a roughly equivalent proportion of the data (98.8%) here. However, as is visually apparent, this is an overestimation for negatively entangled large proteins (our heuristic measure is 7.1 in this region). This example faces a similar issue to the previous, since by uniform sampling we are still considering the scaling of entanglement in relation to number of amino acids.

To strengthen this argument, we will study in more detail the sizes of secondary structure elements as they relate to the size of the entire protein. In Figure 3.15 we plot the average length of SSEs against the length of the chain for each of the proteins in our sample. As can be seen in this scatterplot, there is little increase in the average size of the SSEs for larger proteins. Indeed, the correlation coefficient is 0.1047 with a p-value of 1.14×10^{-24}, indicating that there is strong evidence that there is no correlation between chain length and mean SSE length. In particular, we see a small rise in average SSE length up to chain lengths of 100, but after this point there is no increase. This coincides with Figure 3.12, where the linear bound on entanglement is an overestimation for proteins in this < 100 chain length region. This explains why each of the other backbone smoothing methods we have discussed fail to neatly capture the scaling of entanglement. As we have seen in Chapter 2,

SSEs act like rigid subsections when considered in terms of their contribution to the global entanglement. As a result, the scaling complexity of entanglement is more directly linked to the number of these rigid sections than it is to the number of amino acids in the chain.

3.4 Two topological tangents

To conclude this chapter, we include details of two brief tangential studies to this work. In the first, we investigate the relationship between the knottedness of a protein backbone and its writhe. This work highlights the unique information contained in the writhe of the SKMT smoothed backbones, motivating its use in addition to studying knottedness. The second study similarly uses a unique feature of the writhe perspective, attempting to draw a link between an efficient method of storing cables to the packing of protein backbone curves.

3.4.1 Is there a correlation between knottedness and writhe?

To further investigate the distribution of writhe across the PDB, we consider the locations of a subset of knotted proteins within this distribution. In particular, we consider the set of open-ended trefoil knotted proteins as detailed in [50]. In Figure 3.16 we see the distribution of writhe for this set of proteins in red, compared to the distribution of writhe among our general data set in black. We can see that for larger proteins $L > 35$ there is little correlation between trefoil knottedness and writhe. In contrast, for smaller molecules the presence of a knotted core yields a significant amount of writhing, particularly negative writhing corresponding to left-handed trefoil knotted cores. In the case of longer molecules, the fact that a protein is closer to its bound is often due to the presence of helical geometries, as seen in Figure 3.17. Since it is well understood that random walks are more likely to be knotted than proteins [89], these results make it clear that larger structures often generate systematic complexity through the helical superstructure rather than knottedness. In this way, we see that the writhe measure captures some information that may be lost within the knot based measures.

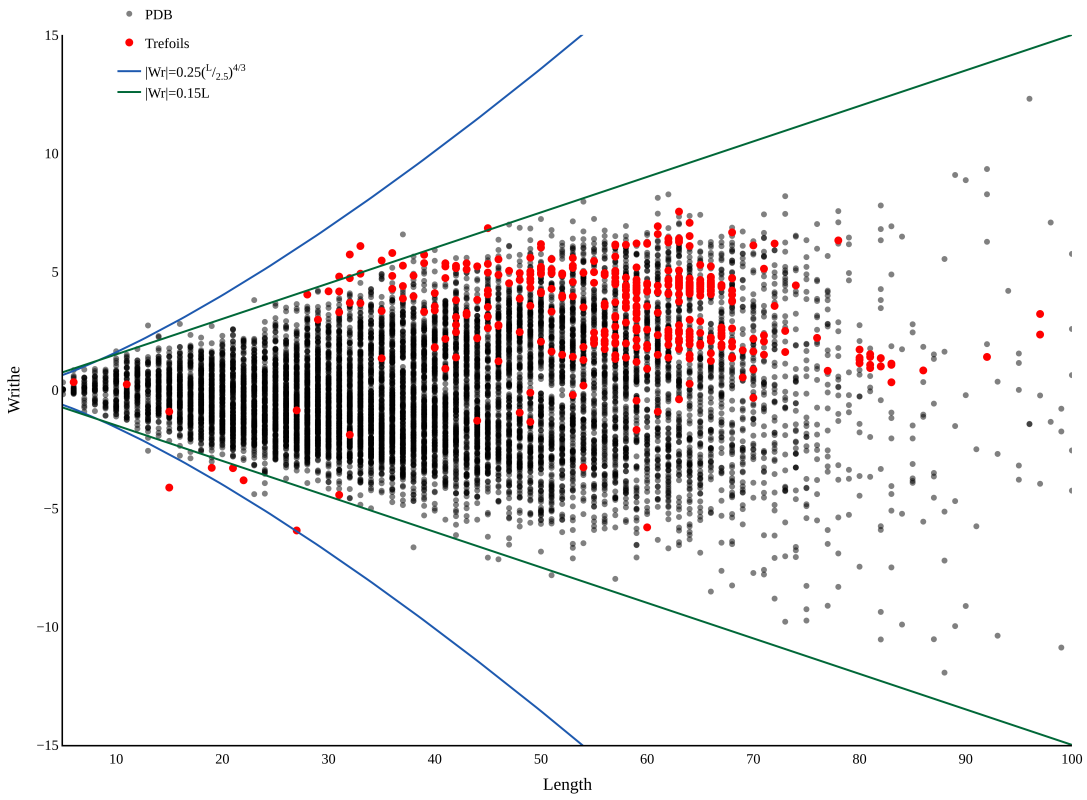


Figure 3.16: The distribution of writhe amongst the open trefoil knotted data set from [50] in red, compared to our subset of the PDB in black.

Figure 3.17: Cartoon representation of a beta/alpha-barrel built by the combination of fragments from different folds (PDB: 3CWO). This protein is an example of a trefoil knotted structure which also has a globally helical structure, thereby maximising writhe.

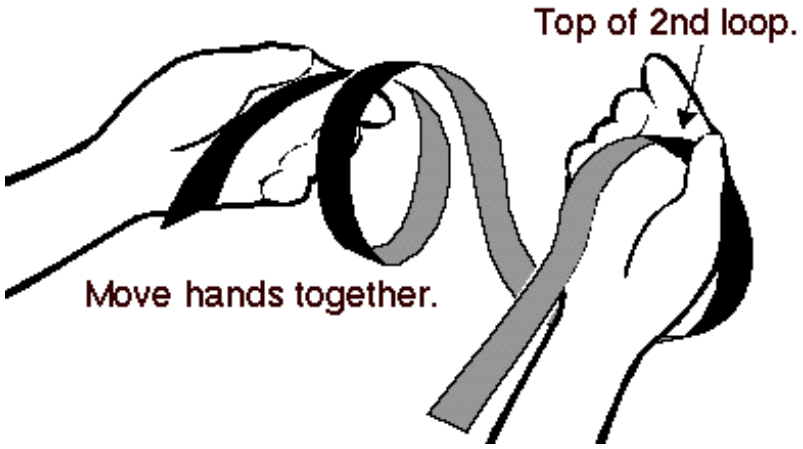
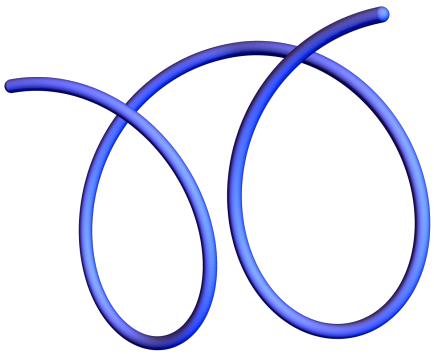


Figure 3.18: How to tie a roadie wrap. https://www.singularsound.com/blogs/news/how-to-wrap-audio-cables

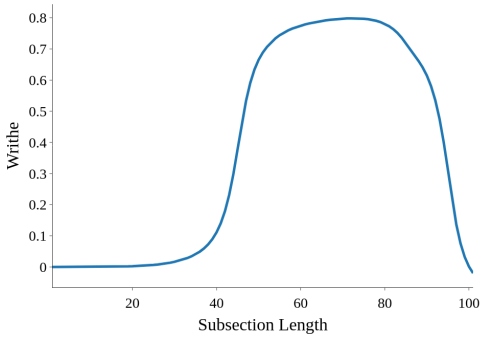
3.4.2 Identifying "roadie-wrap" like structures

To conclude this chapter, we take a brief detour into the music industry and what it teaches us about cable management. Touring the country playing live shows every night comes with an incredible amount of logistical challenges, mostly met by people known as "Roadies". In particular, Roadies have developed a technique to store the many cables and wires a musician needs in a way that prevents knotting and damage, which is known as the Roadie wrap. A schematic diagram of this technique is shown in Figure 3.18. By twisting their hands as they wrap the cable, the orientation of successive helical loops alternates. In terms of writhe, this can be thought of as every positive crossing followed by a negative crossing. As a result, the roadie wrap will have zero total writhe and therefore can be efficiently untied with no risk of entanglement, minimising damage. The structural advantages of tying a cable in this method are clear, so we ask if a roadie wrap like conformation is present as a motif in the folded structure of protein backbones?

In Figure 3.19 we see a Roadie wrap curve along with its writhe profile. The writhe initially increases towards 1, corresponding to the first helical loop, before the cancellation due to the second helical loop, which has the opposite orientation. We will use the characteristics of this writhe profile to define a search criteria for roadie wrap like structures in our protein database. In particular, we will look for subsections of the writhe fingerprint $Wr(\mathcal{C}_{(i,j)})$ which satisfy:



A A smooth roadie wrap curve



B The writhe profile of the roadie wrap curve.

Figure 3.19: An example of a smooth roadie wrap curve with its indicative writhe profile

- $j - i \geq 10$,
- $|Wr(\mathcal{C}_{(1,j)}) - Wr(\mathcal{C}_{(1,i)})| < 0.05$,
- $\exists\, k \in (i, j) \quad \text{s.t.} \quad |Wr(\mathcal{C}_{(1,k)}) - Wr(\mathcal{C}_{(1,i)})| > 0.095.$

Heuristically, we are looking for subsections of the writhe fingerprint where a significant amount of writhe is built up and subsequently cancelled out.

Of the 10742 proteins in our sample, we find that 3543 have at least one subsection of their writhe profile that satisfies the above roadie wrap criteria. In Figure 3.20, we compare the distribution of CATH architectures for this sample with the distribution across the entire data set. We find that there is no significant change in the proportion of architectures represented, suggesting that this super secondary motif is fairly universal.

In Figure 3.22 we present an example of a protein whose backbone contains a roadie wrap conformation, along with its recognisable writhe profile. Although some of the identified roadie motifs exist as subsections of the writhe profile, in this case, essentially the entirety of the backbone forms a roadie wrap conformation.

Given that the roadie-wrap structure is identified via strict writhe criteria, it is a truly systematic method of minimising entanglement within subdomains. Moreover, the physical uses for this structure are well understood in many other contexts. However, the question of whether these functions are applicable to proteins remains

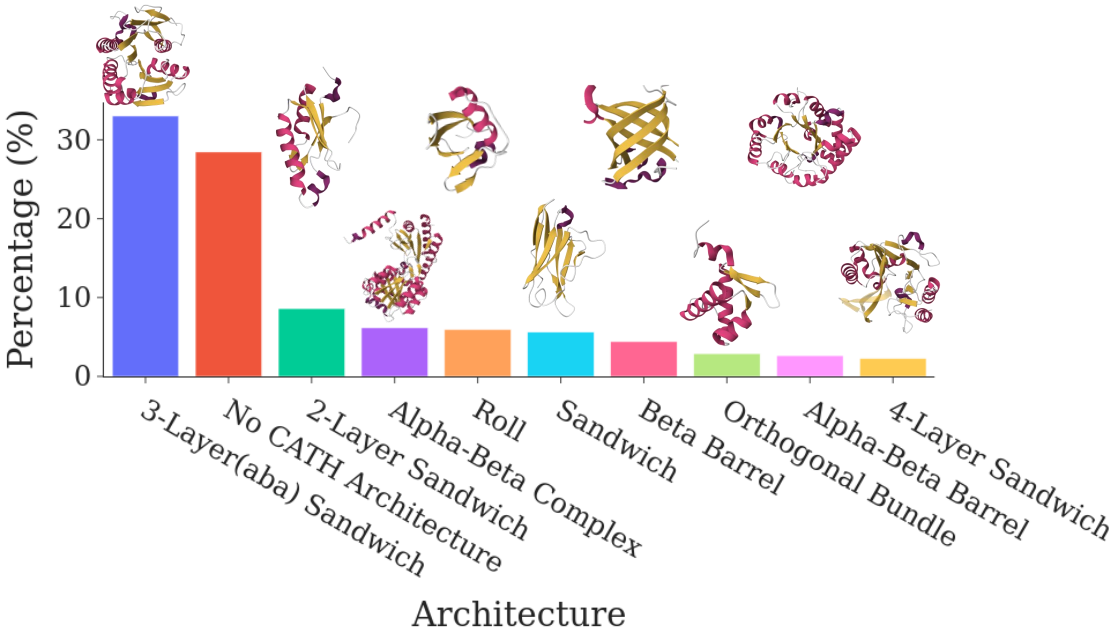


Figure 3.20: Percentage distribution of the top ten CATH Architectures for proteins containing a roadie wrap like subsection. An example cartoon representation of the architecture is shown above each bar.

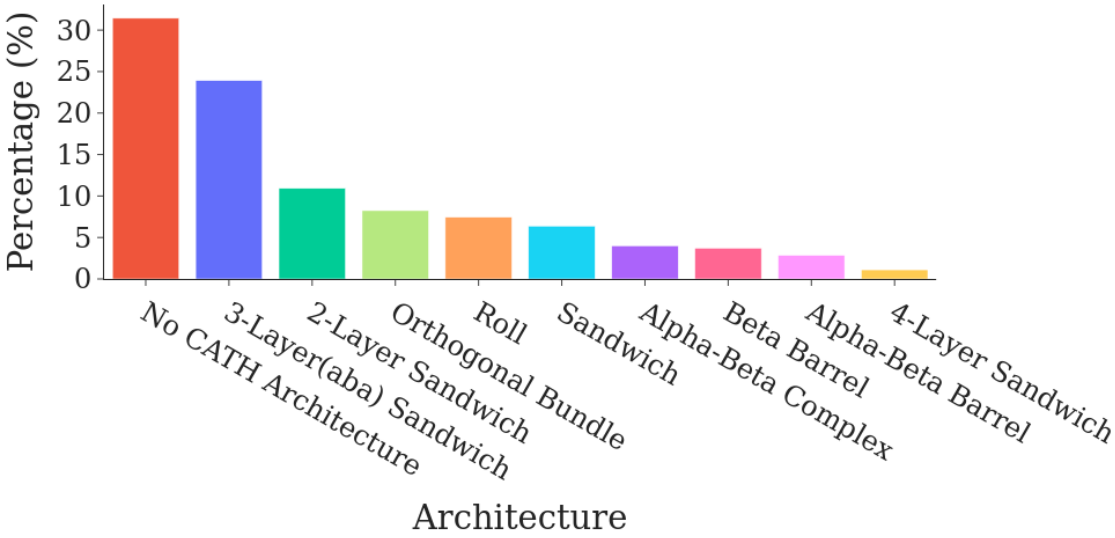
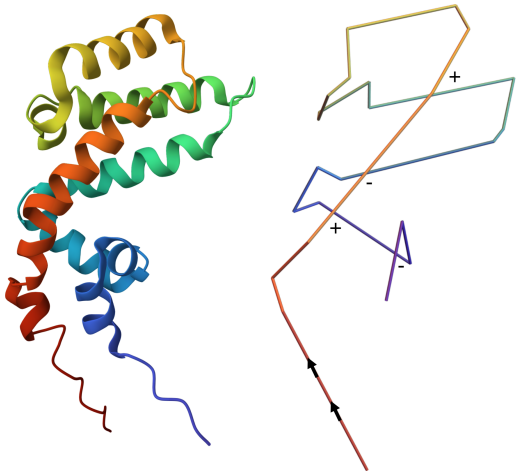
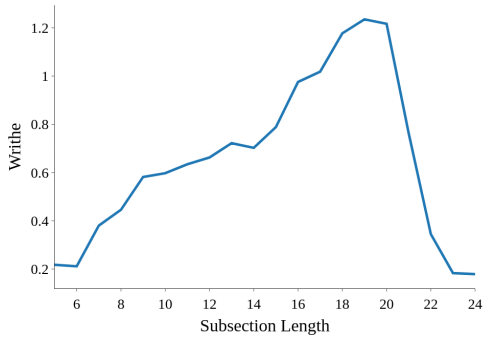


Figure 3.21: Percentage distribution across the whole dataset of the top ten CATH Architectures present in proteins containing a roadie like subsection.



A On the left, the cartoon representation of PDB entry 1DK8. On the right, the SKMT smoothed backbone, with crossings annotated.



B The writhe profile of the SKMT smoothed backbone of PDB entry 1DK8.

Figure 3.22: An example of a protein whose entire backbone forms a roadie like conformation.

open. We present this method for identifying roadie wrap conformations within proteins so that the structural biology community may in the future identify a reason for their prevalence.

We suggest one potential avenue for further study in this area is its relationship to the so-called Greek Key motif. This super secondary motif is defined by the arrangement of four anti parallel β-strands connected by hairpins into a loop which bends back on itself, as seen in Figure 3.23. This motif is commonly observed in proteins, and its 3D structure varies depending on the bonding patterns of the β sheets. The visual similarity between the arrangement of consecutive β-strands into a loop contained within a larger loop of non consecutive β-strands and the arrangement of the helical loops within a roadie wrap is clear.

In Figure 3.24 we see an example of a protein containing a Greek Key motif, along with the SKMT smoothed backbone representation and its writhe profile. Although it may not be immediately visually apparent in the backbone curve, or indeed its SKMT representation, one can see from the writhe profile that this protein does indeed have the build up and cancellation of writhe indicative of a roadie wrap conformation.

A graph-theoretic approach to studying the interactions that may play a role in

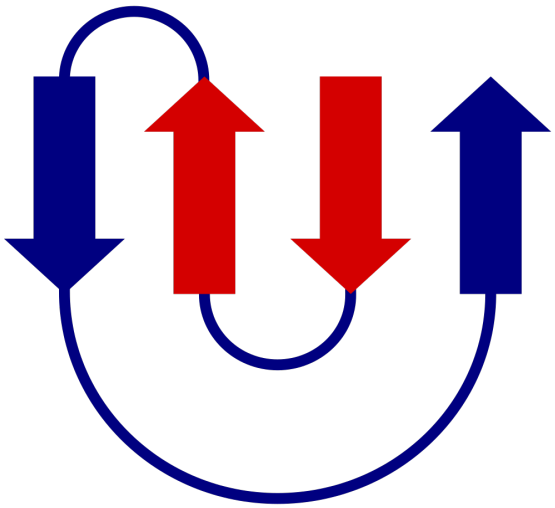
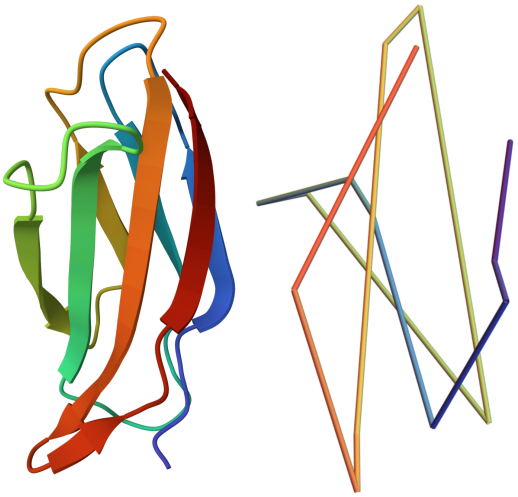
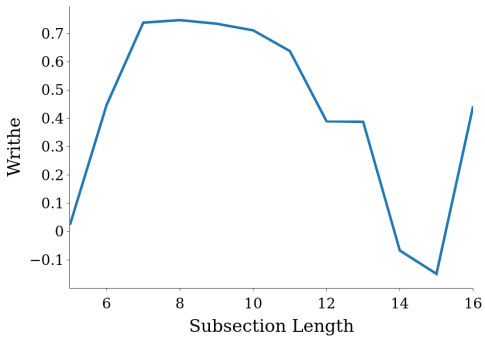


Figure 3.23: A schematic diagram of the Greek Key motif. https://en.wikipedia.org/wiki/Beta_sheet



A On the left, the cartoon representation of PDB entry 1TEN. On the right, the SKMT smoothed backbone.



B The writhe profile of the SKMT smoothed backbone of PDB entry 1TEN.

Figure 3.24: An example of a protein containing a Greek Key motif, whose backbone forms a roadie wrap conformation

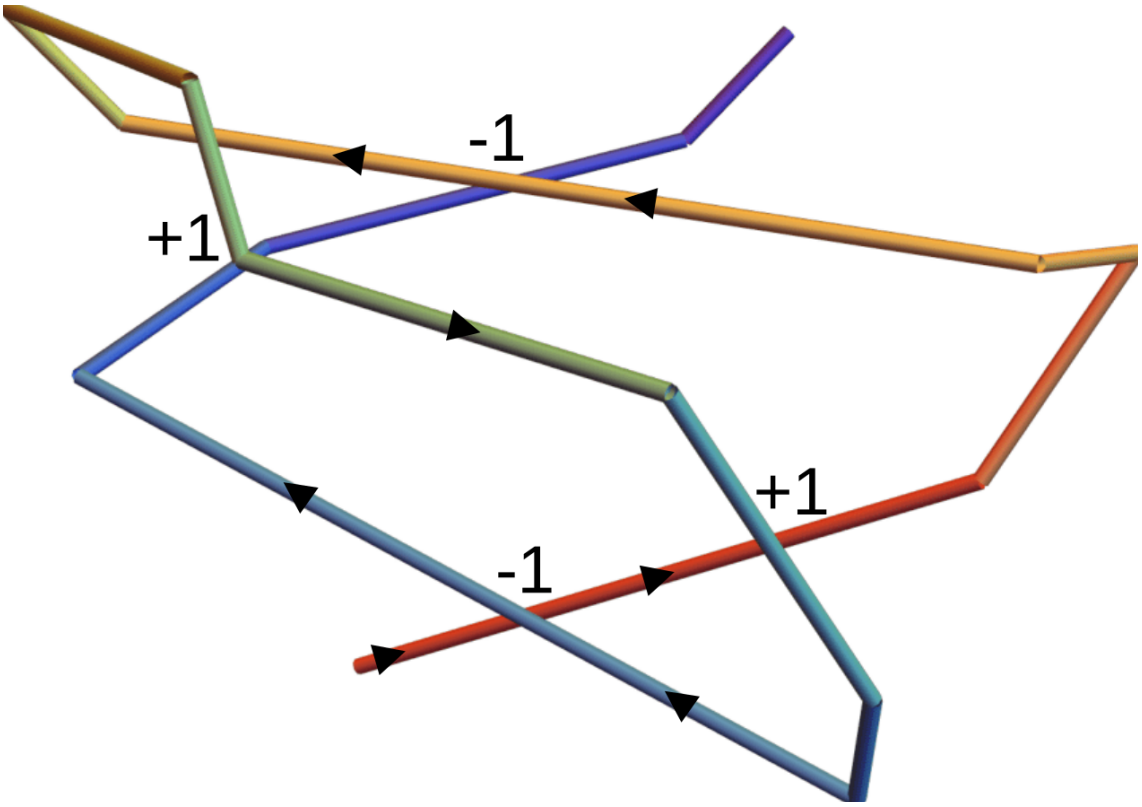


Figure 3.25: The SKMT smoothed representation of PDB entry 1DGN, with significant crossings contributing to the roadie geometry highlighted.

forming the Greek Key topology was presented in [90]. In particular, they considered nine proteins from three different families that each contain Greek-key motifs. We find that a roadie wrap geometry is present in each of the proteins from their dataset, with an example included in Figure 3.25. It has been noted that there can be a lot of variability in the sequences and lengths of the secondary structure elements within Greek Key motifs [91]. The roadie wrap criteria presented here could provide an efficient method to identify a more general version of these common motifs that is independent of the exact secondary structure type.

3.5 Discussion

In this chapter, we presented a novel algorithm for smoothing a protein's backbone curve. The algorithm reduces a backbone curve to a minimal representation of the global entanglement of its secondary structure elements. We studied the distribution of writhe for the smoothed backbone curves and showed that there is a clear scaling of writhe with respect to a sensible choice of length. By investigating this scaling relationship further, we uncovered consistent helical super secondary structure common across many families of proteins. We also discussed the existence of another common super secondary structural motif with a clear systematic entanglement profile.

The smoothing algorithm presented in this chapter integrates some of the most effective elements of other backbone smoothing algorithms in the literature. In particular, it is heavily inspired by the KMT algorithm of [44, 79], which is most commonly used to reduce the computation time of costly knot identification algorithms. Where the KMT algorithm is effective in producing a minimal representation of a curve's topology, it has limitations when applied directly to protein backbone curves. In particular, any recognisable secondary structure will be lost. This preservation of secondary structure arrangement is a key feature of many other backbone smoothing methods in the literature. The approach presented in this chapter, by restricting the KMT algorithm to act on each secondary structure element in sequence, has the ability to provide a minimal representation of topology which preserves secondary

structure arrangement, which is its key advantage over other techniques.

One limitation of this backbone smoothing approach is its dependence on the assignment of the secondary structure to segment the backbone curve. There are many cases where the secondary structure type of amino acids may be mislabelled, or in the case of intrinsically disordered proteins, there may be no recognisable secondary structure at all. We will address the problem of mislabelled secondary structure types in Chapter 4. For the case of disordered subsections, as a first approximation, one could treat them as a long linker and proceed with the algorithm as written. However, their inherent flexibility contradicts the underlying assumption of any smoothing algorithm; that we are working from a static picture of the backbone curve from which we can produce a simpler representation.

By studying the distribution of writhe of a representative sample of protein backbone curves from the PDB, we uncovered a consistent scaling in entanglement complexity with respect to a sensible choice of length. The empirical writhe bounds derived in this chapter show that the range of possible tertiary structures is limited by the secondary structure. The existence of an upper bound on entanglement naturally implies that there are specific arrangements of secondary structures which can maximise writhe. Indeed, one key result of this chapter is the discovery of helical super secondary structures, present across a wide array of proteins. We will see in Chapter 5 that these helical geometries are associated with thermally stable structures with strong inter-sheet bonds.

A common difficulty in studying experimentally determined structures is the potential for missing residues; in particular, the N-terminal residues may be missing from the PDB files. We showed that the empirical writhe bounds determined in this Chapter are robust to this change, further strengthening the argument that the SKMT method is best equipped to represent the global entanglement of the backbone. To extend this argument, we showed that the distributions of writhe for various other methods of backbone smoothing do not have clear scaling in complexity across all length scales.

We concluded this chapter with the inclusion of some exploratory theoretical work, first investigating the links between knottedness of the backbone curve and

the writhe of its SKMT representation. We found that there is no strong correlation, with knottedness contributing little to the global entanglement relative to features such as helical coiling. To further extend this work, we could expand our data set to include different types of knot, or even consider the effect of the depth of the knot, as in [44]. We also investigated the existence of a super-secondary motif which systematically minimises the build-up of writhe. We presented a method for identifying this super-secondary motif and show that it is common across many proteins. There is still much room for further work in this area to uncover the biological explanation for the existence of this conformation, which has many functional advantages in wider contexts.

CHAPTER 4

Carbonara: A rapid method for SAXS-based refinement of protein structures

In this chapter we first introduce the theory underpinning the Carbonara software. This includes the constrained backbone algorithm, and the calculation of the theoretical scattering curves. We then study the distribution of *acn* for a representative sample of proteins from the PDB. This distribution presents a clear lower limit on the expected entanglement with respect to the secondary structure content in the backbone. We will use this lower limit as penalty for unrealistically entangled predictions within the Carbonara software. We will then discuss some further developments and validation of this software. We conclude the chapter by building the foundations of a potential complementary model for Carbonara, and provide a proof-of-concept example of its application. Development of Carbonara has been joint work with Josh Mckeown, with a publication currently under preparation. A significant portion of my PhD was spent preparing Carbonara for installation with my industrial partners Diamond Light Source. This work can be found at https://github.com/mckeownish/carbonara/tree/diamond, with my specific focus being the interactive iPython notebooks for setting up and post processing predictions. This includes methods for preparing the backbone curve for

fitting, identifying a curve's potential secondary structure based on the results in the previous chapter, and developing C++ code to impose the *acn* constraints discussed in this chapter to ensure the algorithm only searches a realistic tertiary fold space. Work on constructing the complementary model was based on private correspondence with Iolo Jones. The complementary model is availabe at https://github.com/arronelab/OptimisedCarbonara

4.1 Carbonara background

4.1.1 The constrained backbone algorithm

The Cα backbone curve is represented by a discrete curve $\mathcal{C} = \{\mathbf{x}_i\}_{i=1}^{n}$. The backbone curve $\mathcal{C}$ is divided into k distinct subsections of length l_k, according to the input secondary structure. At each iteration, the algorithm will construct a new local conformation for a selected (most commonly linker) subsection to produce a new global conformation for the backbone curve. To do this, we take the final three points of the preceding subsection $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, then, given a pair of curvature-torsion values (κ, τ), we solve Equations (2.3) and (2.5) for values θ and ϕ such that the location of $\mathbf{x}_4$ is given by

$$\mathbf{x}_4 = \mathbf{x}_3 + 3.8(\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$$

where $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi]$, and the coefficient 3.8 is the average distance between Cα atoms. We can then repeat this taking the final two points of the previous subsection, with the new first point of the current subsection, to produce the second, and so forth. In order to produce a locally realistic backbone curve model, the distribution of curvature and torsion values along the backbones of a representative sample of proteins from the PDB was computed. Example of these distributions are shown in Figures 2.3 to 2.5, with regions corresponding to the secondary structure types coloured. Further distributions were determined for the geometry of joining subsections, *eg* a linker into an α-helix. The details of the calculation of these probability density functions and the precise formulation of the constrained backbone algorithm can be found in the supporting information of [23].

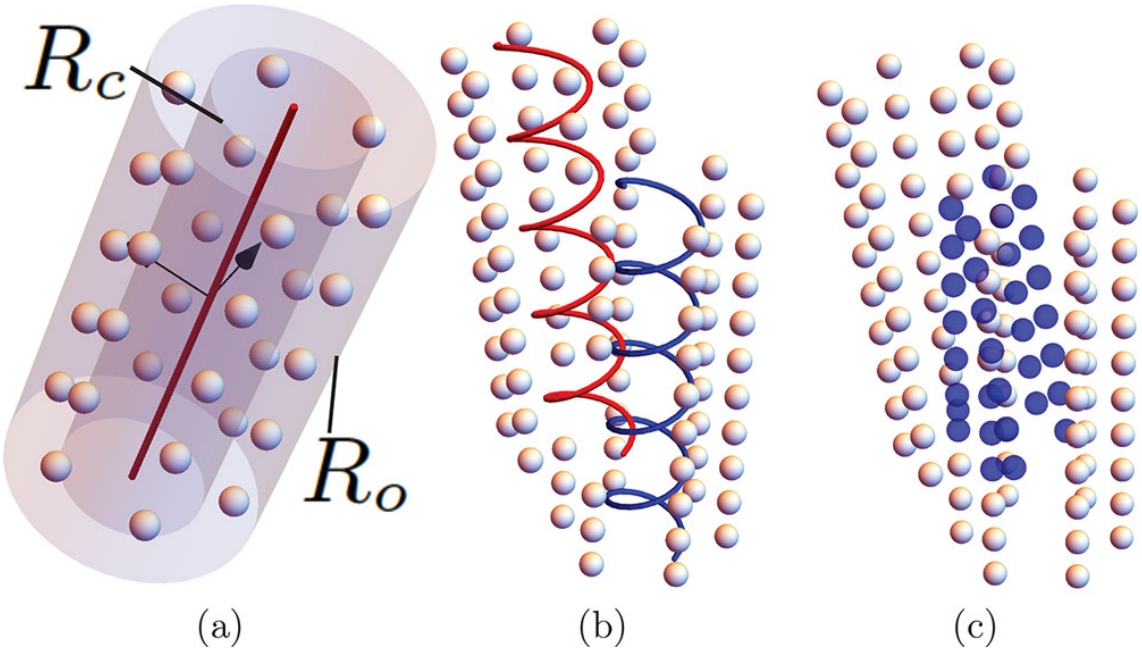


Figure 4.1: An illustration of the hydration shell model from [23]. (a) The backbone curve is surrounded by two cylindrical surfaces which are filled with water molecules. (b) shows two backbone curve subsections who hydration layers overlap, with the water molecules to be removed shown in blue in (c).

4.1.2 The theoretical scattering curve.

In order to compute a realistic scattering from the backbone, it is important to include a hydration layer in the model. This represents the effect of the scattering of the incident X-ray beam due to the water molecules enveloping the backbone in solution with an adjusted density. This is achieved by surrounding the axis of a subsection of the backbone curve with two cylindrical surfaces. Solvent molecules can then be placed between these two surfaces to give an even covering of the backbone curve. For sections where these cylindrical surfaces overlap, solvent molecules are removed. This ensures that solvents deemed too close to the backbone or shared by two sections are not included. Taking into account such inhomogeneities in the hydration layer is crucial to producing an accurate prediction [92]. An illustration of this method is shown in Figure 4.1 reproduced directly from [23].

Once this hydration layer is in place, the scattering intensity is given by the Debye equation [93]

$$I(q) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}}, \tag{4.1}$$

where N is the combined number of Cα backbone and solvent atoms, $q = \pi \sin\theta/\lambda$ is the momentum transfer, and r_{ij} is the distance between atoms i and j. Due to the random motion of particles in solution, we assume that particles take up every possible orientation. Equation (4.1) is therefore a spherical averaging of the scattering intensity due to this assumption. The full derivation of this is included Debye's original paper [93], though it is in German, for an English language derivation see [94]. The functions $f(q)$ are known as form factors, which measure the scattering amplitude due to an isolated atom. Typically, the form factors would be unique to each amino acid, given by

$$f_{am}(q) = f_b(q) - \rho_{ex} f_{ex}(q), \tag{4.2}$$

where f_b represents the scattering due to the amino acid in a vacuum, and f_{ex} is an adjustment according to the excluded volume (that is, the scattering of the volume of bulk solvent that cannot be where the amino acid is). However, an averaged form factor is shown in [23] to keep computation time reasonable whilst maintaining an accurate scattering profile. The averaged form factor is of the form

$$f_b(q) = \sum_{i=1}^{5} A_i e^{-B_i q^2} + C \tag{4.3}$$

with the constants determined empirically. This approach significantly speeds up the calculation time. By separating distances r_{ij} into bins of fixed width and assuming a shared form factor f_b for all amino acids, the contribution to Equation (4.1) from all distances within a given bin can be approximated by a single calculation.

To capture the excluded volume effect of a single atom, an exponential model is also used in the form

$$f_{ex}^a(r_w, q) = v(r_w) e^{-\pi q^2 v(r_w)^{3/2}} \tag{4.4}$$

where

$$v(r_w) = \frac{4\pi}{3} r_w^3 \tag{4.5}$$

is the average atomic radius of the atom common to other SAXS profile methods

(*eg* [18, 19]). For the excluded volume of amino acids, we compute

$$f_{ex}^{am}(q) = \sum_{i=1}^{N_{am}} f_{ex}^{a}(r_w, q) \frac{\sin qr_i^a}{qr_i^a} \tag{4.6}$$

given coordinates for each of the 20 amino acids, and values of r_w for carbon, nitrogen, oxygen, hydrogen, and sulfur. Here, N_{am} is the total number of all such atoms in the amino acid, and r_i^a is the distance for atom i from the central Cα atom. For consistency with the averaged form factor f_b, f_{ex}^{am} is also averaged over all 20 amino acids, weighted according their prevalence in globular proteins [95]. We denote the averaged function by f_{ex}.

The constant ρ_{ex} in Equation (4.2) controls the effect of scattering due to the excluded volume relative to f_b, and is constrained to values within 0.75-1.25 for consistency with previous results [18, 19].

Finally, we compute the scattering due to an individual water molecule from the hydration layer via:

$$f_h(q) = \rho_h(2f_{hy}(q) + f_{ox}(q)). \tag{4.7}$$

Here, f_{hy} and f_{ox} correspond to the vacuum scattering of hydrogen and oxygen respectively, and ρ_h is an empirically determined constant which represents an adjustment to the scattering amplitude. This adjustment is due to the hydration effect, which is the observed difference between the density of the bulk solvent and the density of the hydration shell (see *eg* [96]).

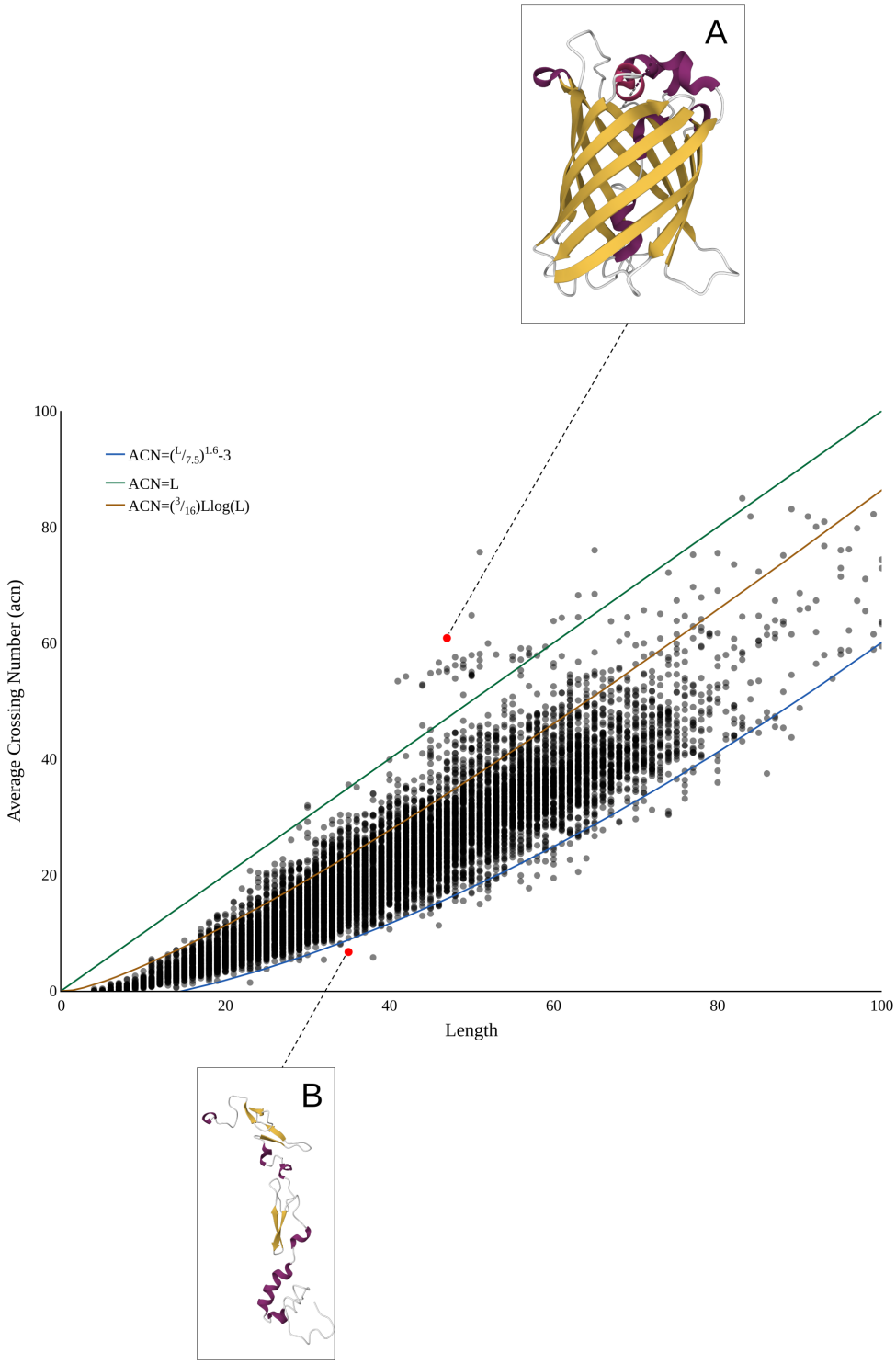


Figure 4.2: The distribution of *acn* for the SKMT smoothed backbones of a representative sample of $> 10,000$ proteins. In blue, an empirically determined lower bounding curve. In orange, the $O(L \log L)$ growth in *acn* as in [73]. In green, linear growth in *acn* with respect to length. Inset: A: PDB Entry 3EVP. B: PDB Entry 1DAN.

4.2 Limiting Carbonara to searching a realistic tertiary fold space

4.2.1 Length constraints on the absolute complexity of entanglement of proteins.

Although the predictions of the backbone curve produced by the model in [23] are guaranteed to be locally realistic due to the curvature torsion constraints, there was no control over the plausibility of the global conformation. We have seen in the previous chapter that there is a limit to the range of possible entanglements relative to the secondary structure content. To include such a constraint into Carbonara, we aim to find a similar bounds on the absolute entanglement as measured by the *acn*. We plot the distribution of *acn* for our representative sample of SKMT smoothed protein backbones in Figure 4.2, and will discuss some of the key features of this distribution.

1. 99.5% are bound from above by a linear growth of complexity $acn(\mathcal{C}) = L$.
2. 98.9% have an *acn* measure above the curve $(L/6)^{1.6} - 3$, a fit obtained by sight.
3. A curve $(3/16)L\log(L)$ also acts as a reasonable upper bound, with 91.4% of the data falling below this curve.

The first point is commensurate with the conclusions of the previous chapters. That is, the complexity of entanglement in protein tertiary structure arises from large scale helical geometries. Similarly to the distribution of writhe, we see that some larger proteins can see superlinear growth in their *acn*. In particular we see a cluster of super-linear proteins with L between 40-66, which we will discuss further in Chapter 5 using a writhe based similarity metric which will be introduced in that chapter.

The second point implies a possible lower limit on the amount of complexity with respect to secondary structure. We will assess the strength of this lower bounding

curve in the following subsection by investigating those protein's whose *acn* falls below. Later, in Section 4.2.3, we will see how we can use this empirically determined lower limit as a penalty to limit Carbonara to producing predictions that have a globally realistic conformation.

On the third point, in [73] it is shown that $acn(\mathcal{C})$ for an equilateral random walk $\mathcal{C}$ of length n grows like $nlog(n)$. This bound is further studied in the context of proteins in [74] where the *acn* of a sample of proteins is shown to follow the same length scaling law. The coefficient of 3/16 determined in [74] is seen to fit as a upper bound for small proteins up to length 15. This is mostly coincidental however, given our specific definition of length and the fact that the SKMT smoothed backbone is clearly not equilateral, we do expect quite a different distribution of *acn* in our context. In particular, the long rigid secondary structures could act as a barrier to entanglement, unless arranged in such a systematic way as to maximize entanglement, as in the case of the many helical structures. For example, the signaling protein seen in Figure 4.2A (a representative of the group of proteins whose *acn* exceeds the growth of $O(L)$) forms a large β-barrel structure. It is clear that a uniform random walk would be highly unlikely to achieve this specific helical and symmetric structure.

4.2.2 Investigating the outliers to the lower bounding curve

In order to use the empirically determined bound in a predictive capacity, it is important to understand those proteins whose *acn* lies below this curve. Figure 4.2B is an illustrative example of the majority these cases. Namely, it is a relatively trivially entangled sub unit of a more complex multimeric structure. Given this study was designed to understand the self-entanglement of monomers, for any PDB file that contained multiple chains, we simply extracted the first chain for analysis. We find that 63.6% of the points whose *acn* falls below the empirical lower bound are due to this treatment of multimers during data scraping.

For the remaining 36.4%, we find that their position is due to a poor assignment of the secondary structure that affects the SKMT curve. In particular, we found occurrences of single amino acid long α-helices and β-strands. To determine a con-

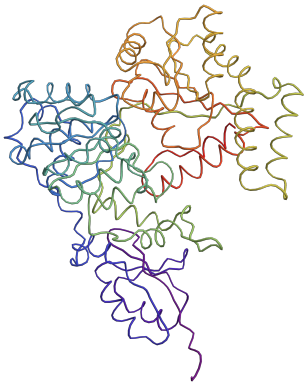
Figure 4.3: An example of a backbone curve with poor secondary structure assignment. α-helices are shown in red and β-strands in yellow. Note in particular the single residue long β-strands in the top left and bottom right of the figure.

sistent helical geometry requires at least three points, so we find these single residue secondary structures to be implausible. An example of this is shown in Figure 4.3, where many single residue long β-strands are present. Given the relationship between the length of the SKMT curve and the number of distinct SSEs, the inclusion of single residue SSEs means that these structures are misplaced in the *acn* distribution. Though we included a routine to identify obvious mislabelling in the initial data scraping, there remains some examples such as this one where their SKMT length is affected.

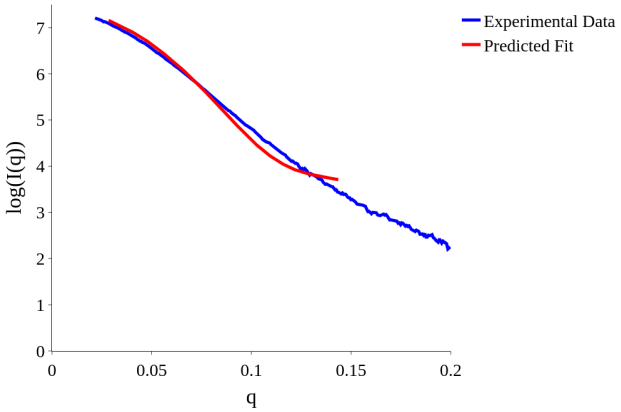
By clarifying these outliers, we are confident that we can safely use the lower bounding curve from Figure 4.2 as a penalty in Carbonara for unrealistically entangled monomer units. Indeed, secondary structure cleaning techniques are included in Carbonara as part of the prediction set up.

4.2.3 Human SMARCAL1 - the entanglement penalty in action

We now highlight a concrete example of the effectiveness of the empirical *acn* lower bound for structural prediction. For this, we consider the gene regulatory protein Human SMARCAL1. This protein regulates gene transcription by altering the structure of chromatin around those genes [97]. There is a predicted structure



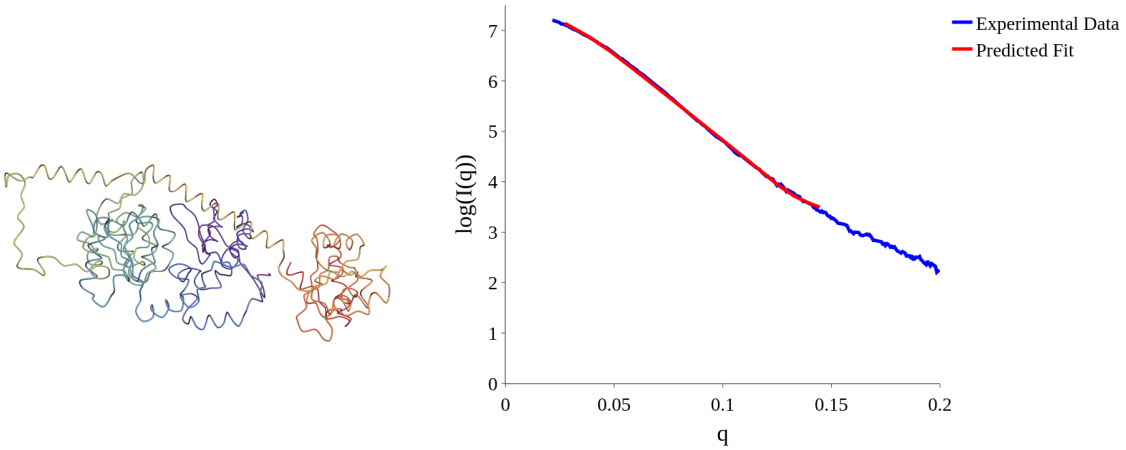
A The AlphaFold predicted structure for Human SMARCAL1.



B The scattering profile of the AlphaFold predicted structure against the experimental scattering data.

Figure 4.4: The AlphaFold predicted structure for Human SMARCAL1 has regions of low confidence, and the fit to the scattering data suggests it opens out in solution.

from AlphaFold for this protein, seen in Figure 4.4A, however it has regions of low confidence, and most importantly is a poor fit to the experimental BioSAXS data (obtained on the B21 Beamline at the Diamond Light Source). This is illustrated in Figure 4.4B where the SAXS scattering model is obtained using the method described in [23] (similarly poor quality fits were obtained using the FOXS web server [19] as a check). When fitting structures to BioSAXS data, it is important to note the different resolution scales corresponding to the q values. In particular, the lower q range corresponds to larger scale structural information with the resolution increasing with higher q. The fitting of a low q range is therefore of paramount importance as a small discrepancy with the data there can mean the overall shape of the molecule is wrong. In contrast, at higher q discrepancies in a fixed prediction are less meaningful. Thus, for this illustrative example we stick to fitting the structure to the $q \in [0, 0.15]$ range. A rough rule of thumb for these experiments is that globular conformations have a hill like shape in their SAXS curve, whilst elongated rod or pill-like structures will have a much flatter curve (see *eg* Figure 3 of [98]). With this in mind, Figure 4.4 indicates that the AlphaFold prediction is too globular compared to the experimental data, suggesting the protein likely opens out somewhat in solution.



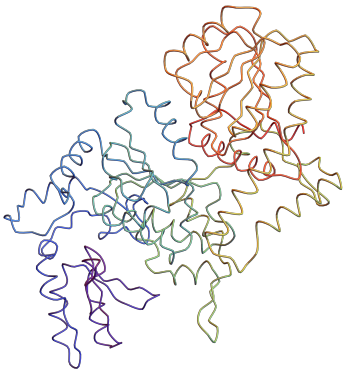
A A prediction of SMARCAL1 which falls below the *acn* bound. ($acn = 48.3$ to 3 $s.f$)

B The scattering profile of the below *acn* empirical bound predicted structure against the experimental scattering data.

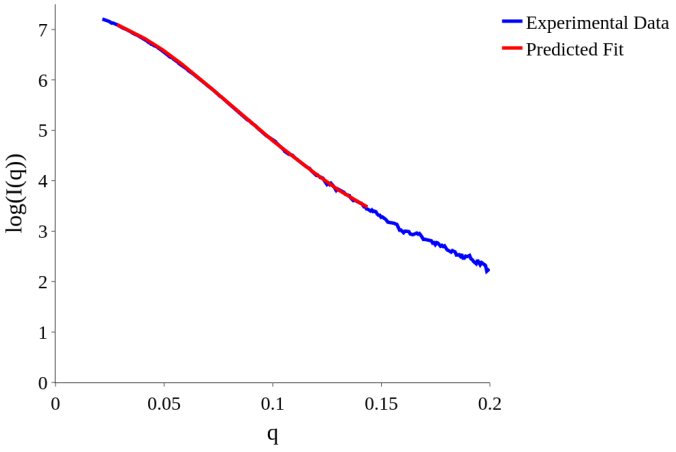
Figure 4.5: Examples of potential backbone structures which fit the SMARCAL1 data very well but are unrealistically unfolded according to the empirical bound on *acn*.

Using the constrained backbone algorithm, we can rapidly produce new potential structures which improve the fit to the BioSAXS data. However, this original method had no constraint on the plausibility of the overall tertiary structure. This is a potential issue as the inverse scattering problem is not well posed [16] and multiple differing predictions for the data can be made. Though a prediction could theoretically be tested using an all-atomistic physics model, this is an extremely time-consuming and intensive process. By contrast, the *acn* calculations are of lower complexity than the scattering calculation itself, and can at the very least be used to rule out good fits to the scattering data which are unrealistically unfolded. We ran a series of fits to the SAXS data on $q \in [0, 0.15]$ and calculated the *acn* of their final prediction.

An example of these fits is shown in Figure 4.5A, with the protein opening out quite significantly. Visually we can see that the three key domains of the structure are too far apart. Since SMARCAL1 has 84 secondary structure elements, we would expect the *acn* of its SKMT curve to be at least 56.8 (to 3 sf) and likely above this. For reference, the original (too globular) structure has an SKMT representation with an *acn* of 69.5 (to 3 sf). This opened out prediction however has an SKMT



A A prediction of SMARCAL1 which has *acn* above the empirical bound ($acn = 59.2$ to 3 $s.f$)



B The scattering profile of the realistically entangled structure against the experimental scattering data.

Figure 4.6: A predicted structure for SMARCAL1 which is both a good fit to the scattering data and is realistically folded according to the empirical *acn* bound

representation with *acn* of 48.3 (to 3 $s.f$). By modifying the constrained backbone algorithm to include a step function penalty for structures whose *acn* falls below the lower bounding curve, we can produce predictions such as the structure seen in Figure 4.6A. This structure is much more globular than Figure 4.5B, although less so than the AlphaFold predicted structure as expected from the shape of the experimental scattering data, and the *acn* of its SKMT curve is 59.2 (to 3 sf), comfortably above the bound. Although this alone is not sufficient to make a complete prediction that the outcome is a plausible structure for the protein; as part of the Carbonara pipeline we can pass these predictions into MD simulations for further analysis. This example highlights the potential efficacy of the measure *acn* as a means of providing a computationally efficient additional constraint on the search space of tertiary folds.

4.2.4 Using Carbonara to optimise low confidence predictions from AlphaFold.

As the landscape of protein structure prediction became dominated by AlphaFold, the focus of Carbonara has shifted away from ab-initio prediction (as it was in [23]).

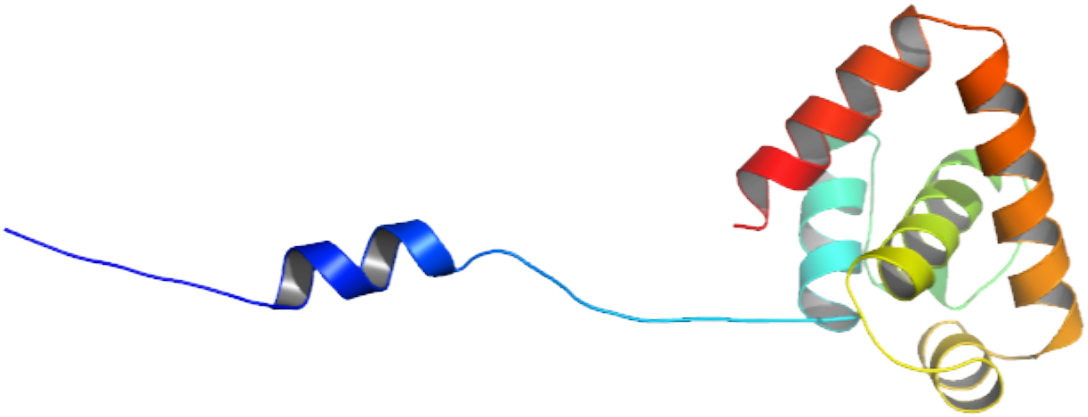


Figure 4.7: The AlphaFold predicted model for domain IV of DnaA.

Now, researchers often have an AlphaFold predicted structure which does not agree with their experimental BioSAXS data. One such example where Carbonara has provided additional insight is in the study of domain IV of DnaA from Bacillus subtilis (UniProt: P05648). By initially unwinding the DNA at the chromosomal replication origin, DnaA allows DNA-replication machinery to load onto single-stranded DNA. Domain IV is highly conserved and essential for this DNA replication, in particular in its role binding to dsDNA. Collaborators at Newcastle University are interested in targeting this interaction for an early-stage small-molecule drug discovery project, screening fragments against domain IV and assessing their interference with this ds-DNA interaction. However, this screening does not provide a high-resolution model of how fragments bind. An AlphaFold model was produced to study this docking and SAXS experiments were performed to quality-check this model.

Figure 4.7 shows the the AlphaFold predicted model for domain IV of DnaA, with its fit to the experimental scattering data seen in Figure 4.8. The poor fit is mainly due to the opened out nature of the structure with the N-terminal linker (in blue) pointing directly away from the rest of the main globular domain.

In Figure 4.9 we see a Carbonara predicted model for domain IV for DnaA, with its fit to the experimental scattering data in Figure 4.10. The N-terminal linker has been refolded to sit much closer to the rest of the structure, whilst maintaining the relative orientation of the C-terminal α-helices. The resulting improvement in the scattering fit is clearly visible.

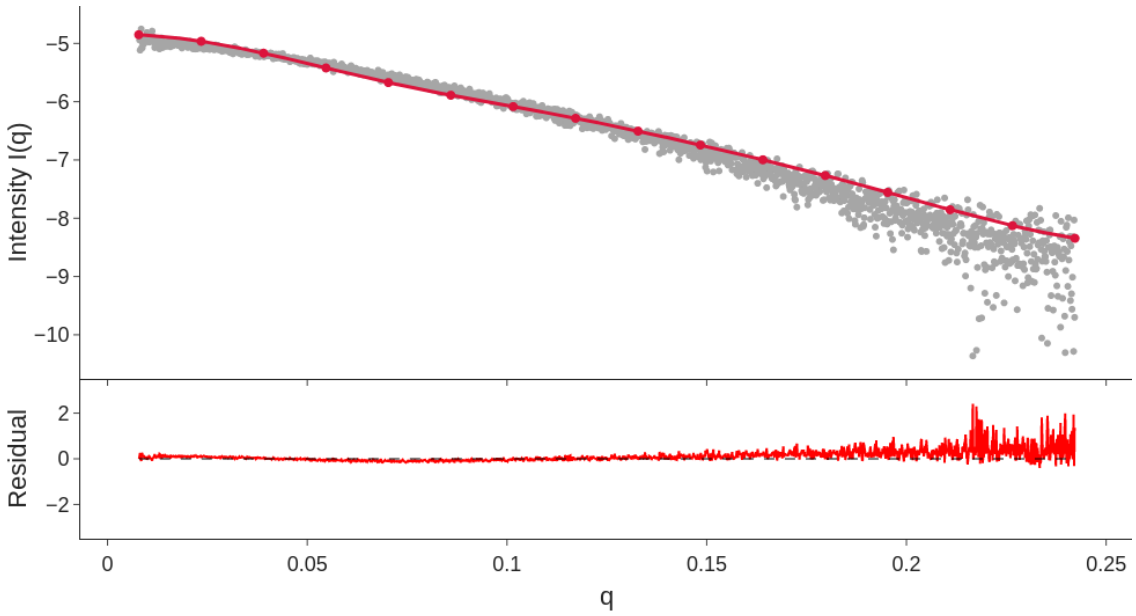


Figure 4.8: In black, the experimental scattering data for domain IV of DnaA. In red, the scattering profile for the AlphaFold model.

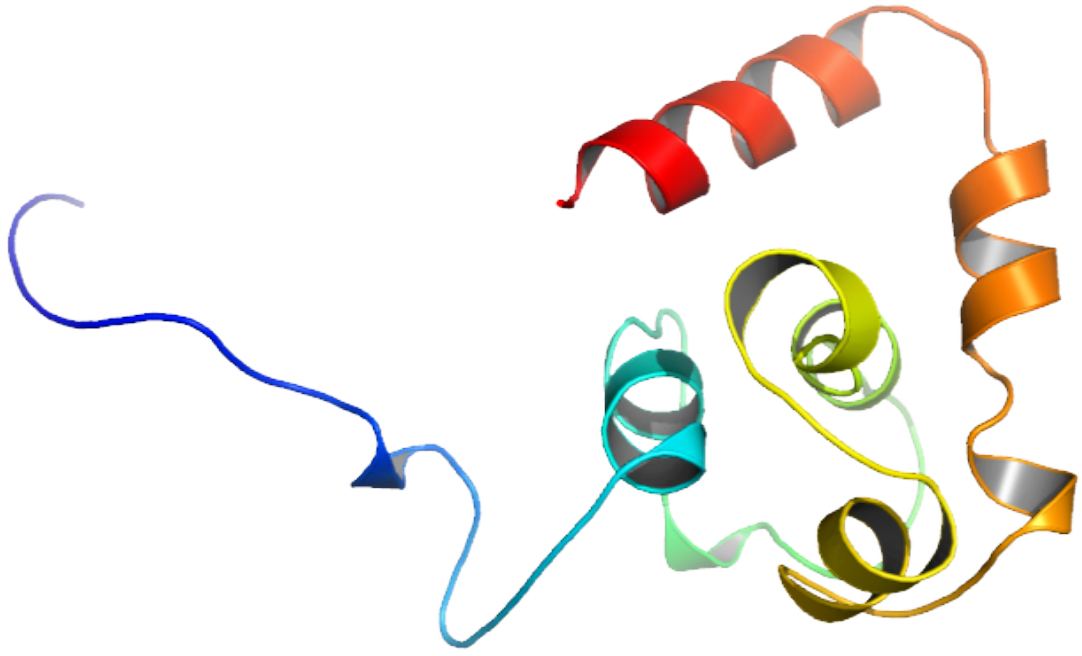


Figure 4.9: A Carbonara predicted model for domain IV of DnaA.

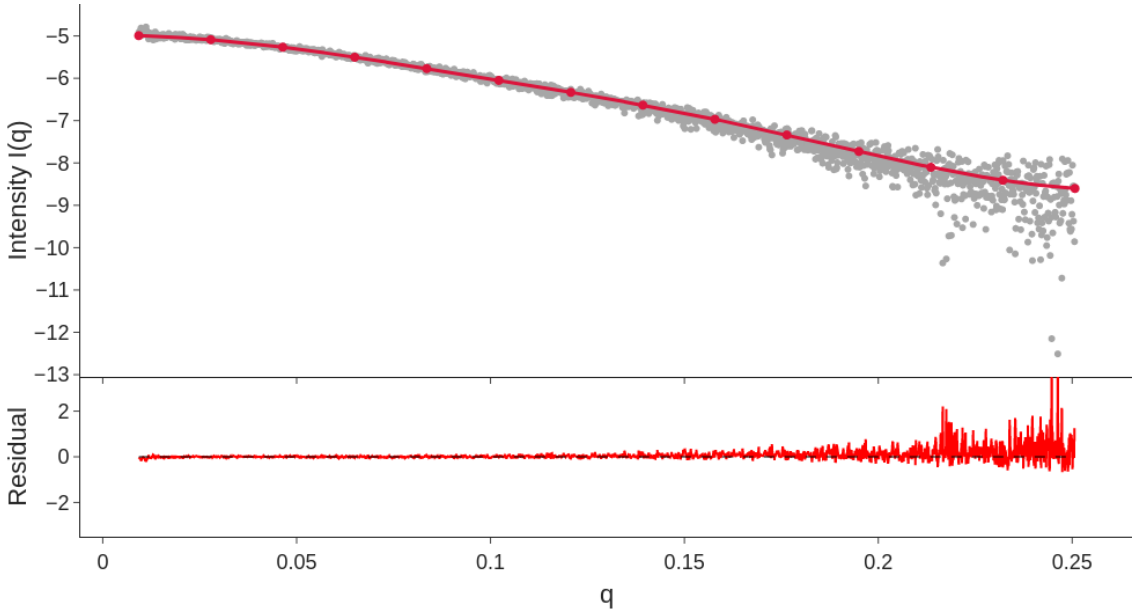


Figure 4.10: In black, the experimental scattering data for domain IV of DnaA. In red, the scattering profile for the Carbonara predicted model.

In fact, there are many Carbonara predicted models which improved the scattering fit, with a large amount of mutual similarity. The extent of the variability across these predictions is shown in Figure 4.11, with 10 predictions aligned and overlaid. The N-terminal linker is the main difference between each of these predictions, indicating that this region is potentially dynamic in solution. With the majority of AlphaFold's training set being crystallographic data, it produces low confidence predictions for dynamic regions such as this.

To study this dynamic nature further, we can use one of the other novel features of Carbonara. Namely, Carbonara can make predictions for scattering data which may be an average of multiple structures in solution, or, as in this case, a protein which itself adopts multiple conformations in solution. In Figure 4.12 we see the fit to the experimental scattering data for a combination of two Carbonara predicted structures. This is a further improvement on the fit for a single state.

In Figure 4.13 we see the two predicted conformations from Carbonara, aligned to highlight the variation of the N-terminal linker between them. Carbonara produces the scattering fit above via a ratio of 30% of the conformation shown in blue and 70% of the the conformation shown in orange. This example highlights the extra insight that Carbonara predictions can provide over AlphaFold models, in particular

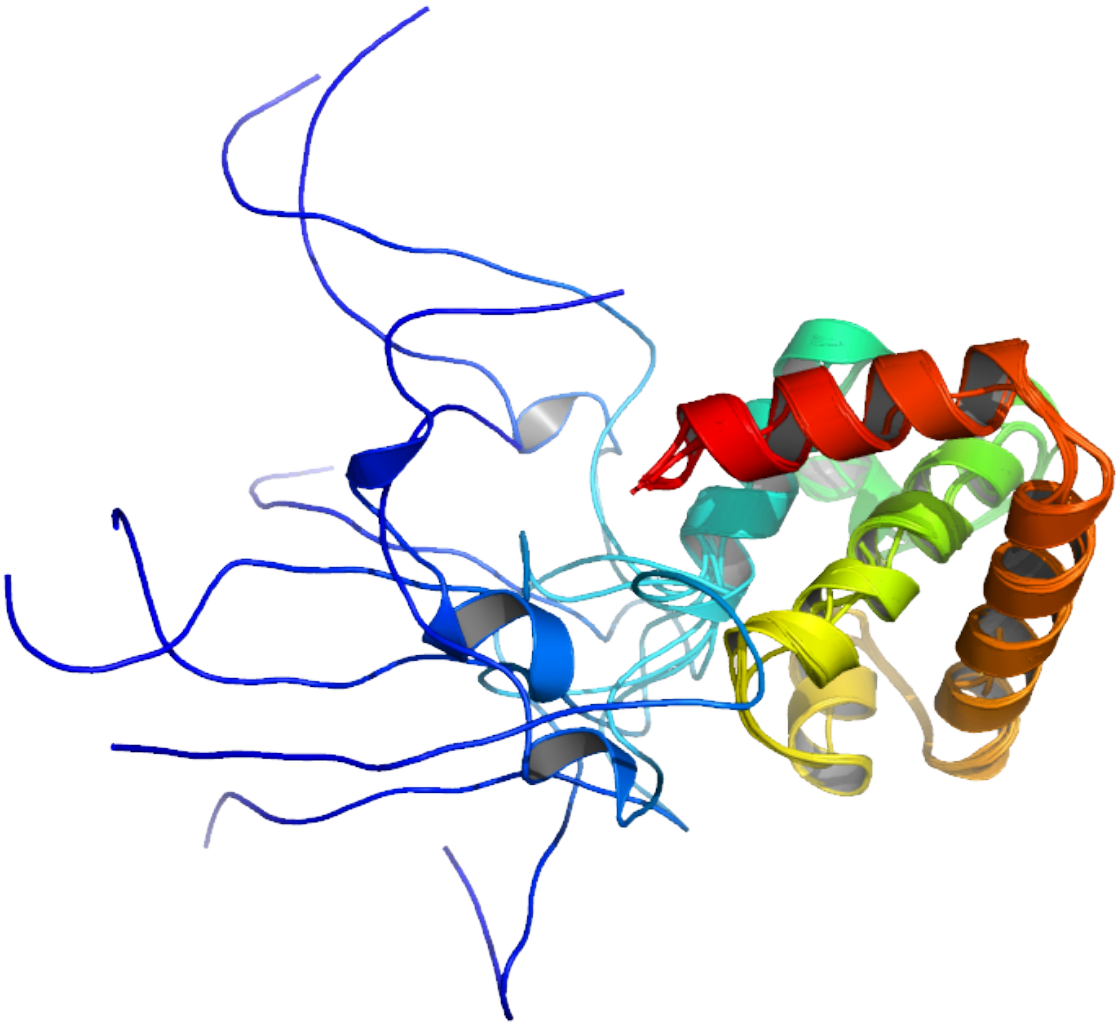


Figure 4.11: 10 Carbonara predicted models for domain IV of DnaA.

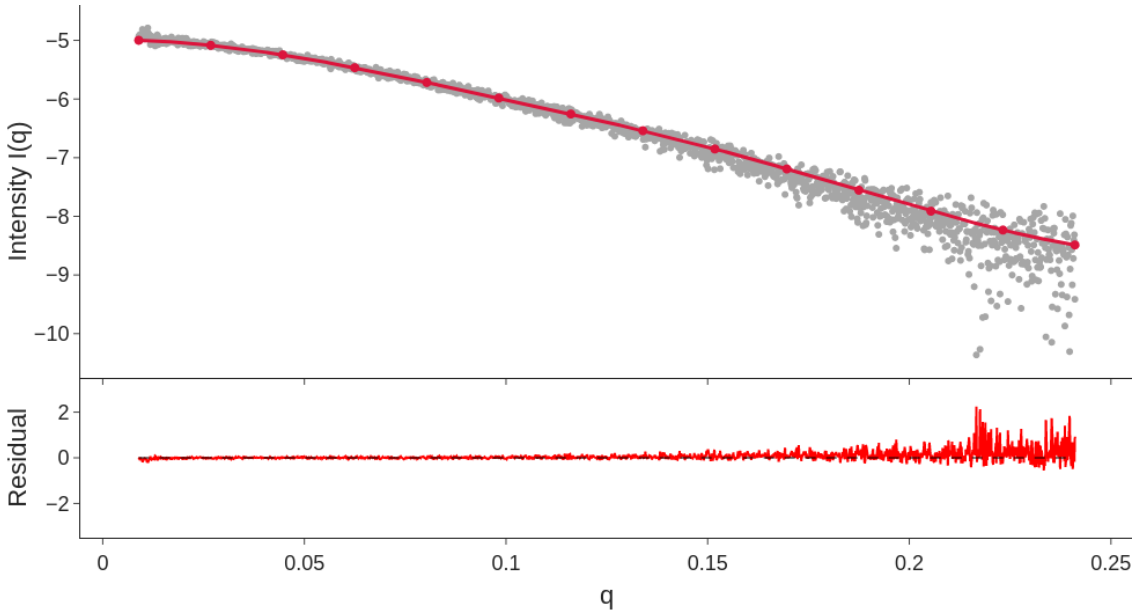


Figure 4.12: The fit to the experimental scattering data for a combination of two Carbonara predicted models.

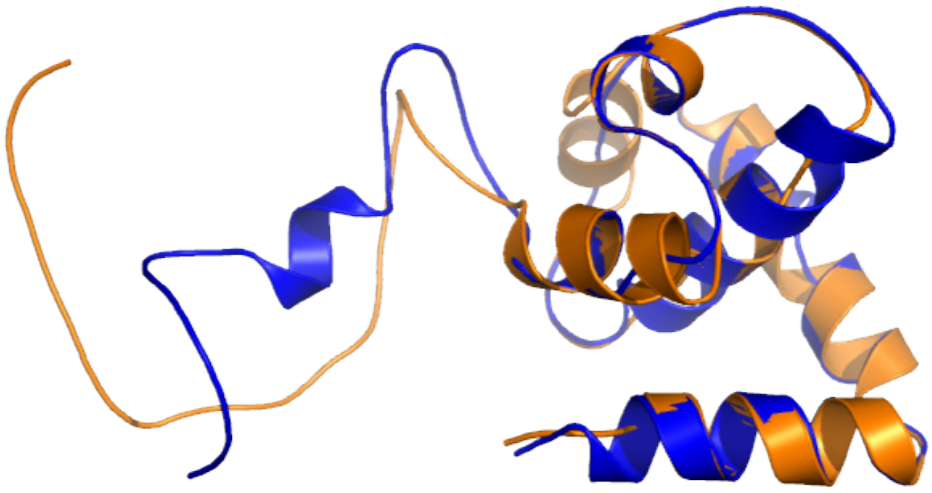


Figure 4.13: Two Carbonara predicted models for domain IV of DnaA

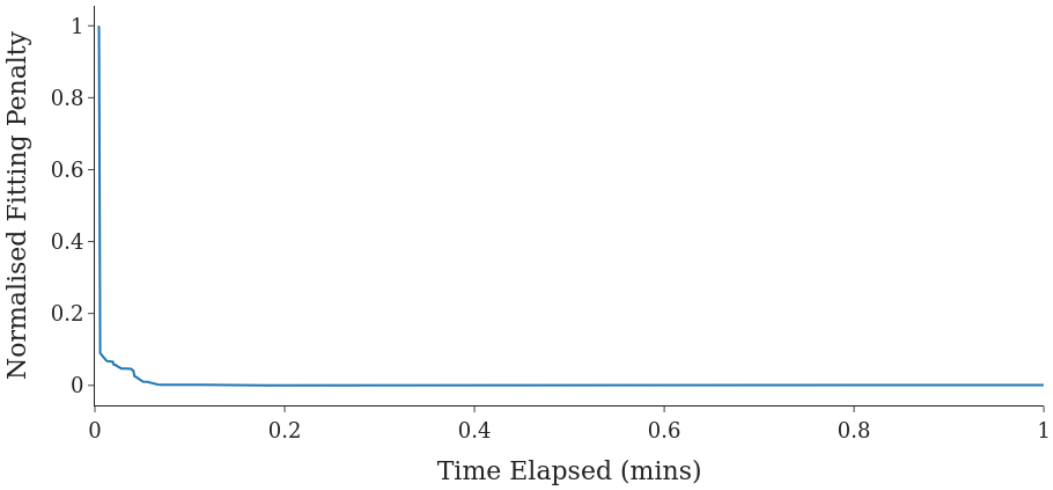


Figure 4.14: An example of the drop off of the fitting penalty over the course of a prediction.

for highlighting potentially dynamic regions in solution. The ability to fit multiple predictions to the scattering data is a key development in this regard.

4.3 Developing a potential complementary model to Carbonara

4.3.1 Motivation

The main advantage of the Carbonara method is its ability to quickly sample many regions of the potential state space, modelling conformations that may be inaccessible to methods such as Molecular Dynamics (MD) simulations. However, this approach does come with a cost. As can be seen in Figure 4.14, the model is able to quickly make large scale changes to the structure to drastically improve the fitting penalty. After this initial improvement however, it becomes increasingly difficult for the Monte Carlo approach to make sufficiently subtle changes to improve the fit. The current solution to this problem is to pass the predicted structures into an MD simulation to allow them to find an energetically favourable state. This remains a costly solution, and we therefore look for a model which could bridge the gap

between these two approaches. We envision a pipeline where Carbonara provides the large scale changes as a first approximation, then this complementary method optimises this prediction, before finally passing into an MD simulation.

4.3.2 A broad description of the approach

At each iteration, Carbonara selects a subset of the linkers of the backbone curve and randomly samples new values from their curvature-torsion distributions to redraw these subsections. Roughly speaking, we sample the curvature-torsion space to produce a change in the Cartesian coordinate space of the curve. Changing the geometry of just one linker in this way can make significant changes to the tertiary structure, *eg* through hinging open two domains. An example of the large scale conformational changes Carbonara makes is shown in Figure 4.15, with the N-terminal domain hinging dramatically around the rest of the molecule. This can make it difficult to make changes to the tertiary structure in a subtler way which could improve the fit to the scattering data of an already tightly folded protein while avoiding overlap.

We will look to build a complementary model that acts in the "opposite" direction. We will take a gradient descent approach, making changes to the coordinates of the curve which optimises the fit to the curvature-torsion distributions amongst other constraints. We will mirror as closely as possible the constraints present in the Carbonara model, namely:

- Cα neighbour distance penalty.
- An overlap penalty for non-neighbouring Cα atoms.
- A curvature-torsion space penalty.
- *acn* penalties for realistic global and subdomain entanglement.

By formulating each of these penalties as probability distributions, we can use the PyTorch optimisation package to perform a gradient descent optimisation of the Cα backbone curve based on its likelihood relative to these distributions. In this manner, this complementary model should make the subtle changes to the tertiary

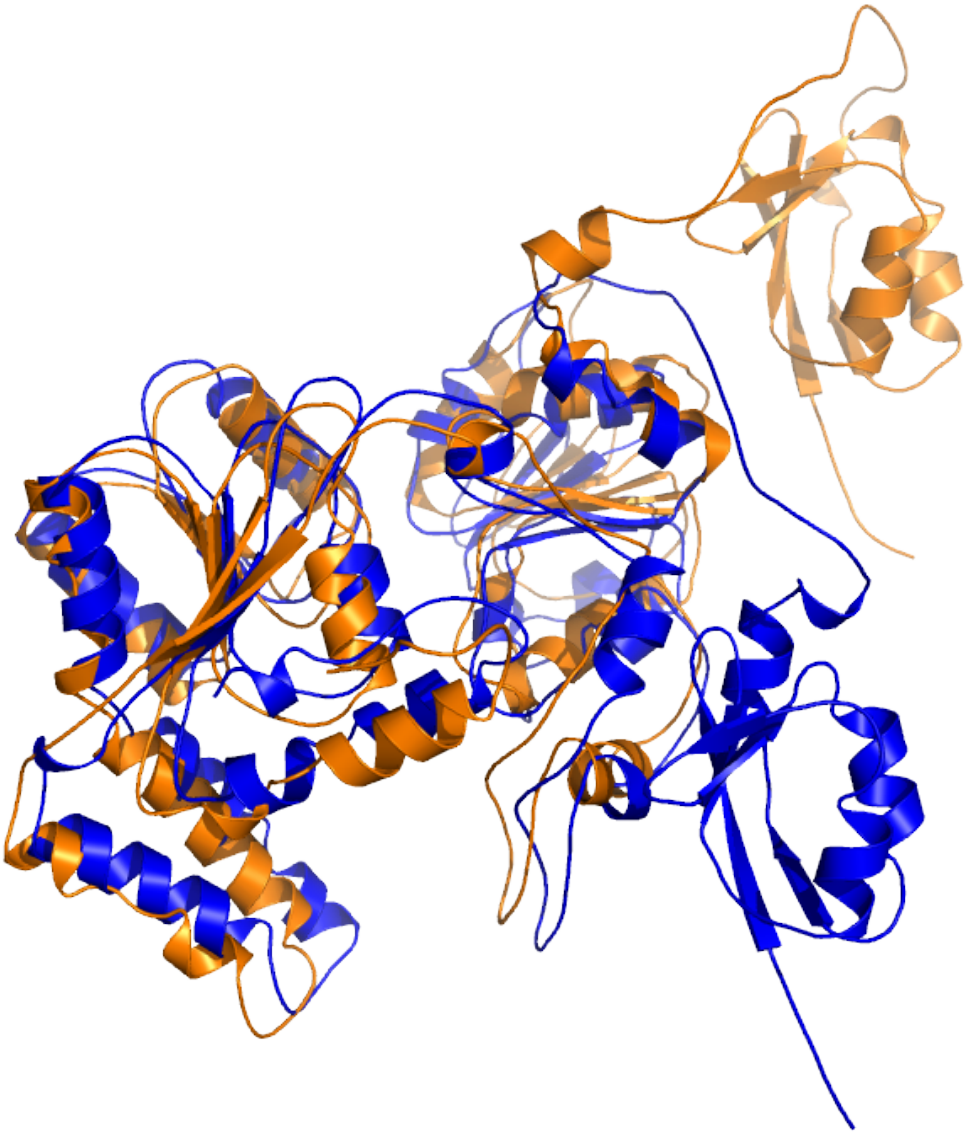


Figure 4.15: The AlphaFold predicted structure for Human SMARCAL1 in blue and a Carbonara predicted structure in orange. Aligned to highlight the large scale conformational changes Carbonara can produce.

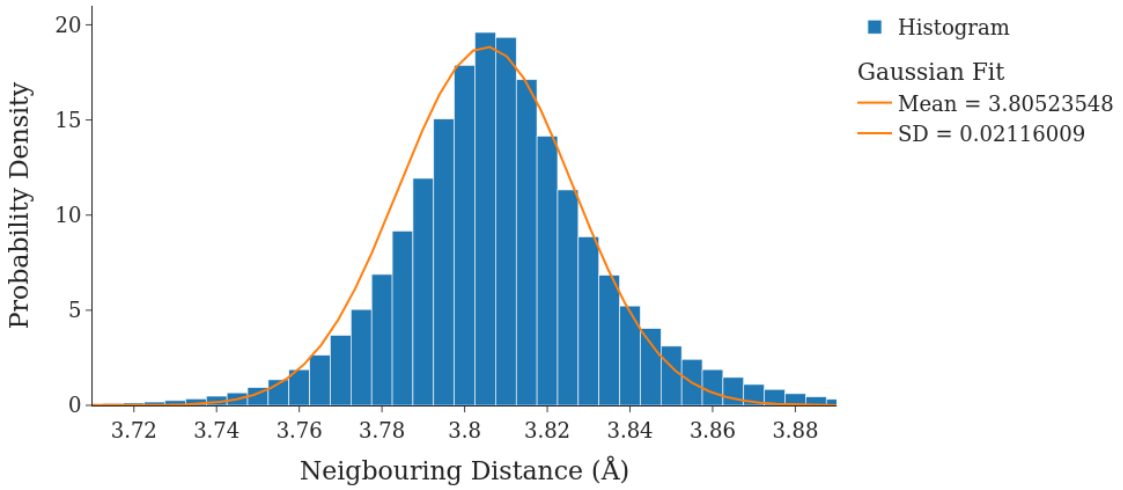


Figure 4.16: In blue, the histogram of neighbouring Cα distances for a representative sample of protein backbone curves. In orange, a Gaussian curve fit to the histogram.

structure which are difficult for the Monte Carlo approach taken in the original Carbonara model. We now detail the formulation of each of these penalties.

4.3.3 The neighbouring Cα distance penalty.

The neighbouring amino acids in a protein are connected by peptide bonds, such that the average distance between neighbouring Cα atoms is 3.8Å (eg [99, 100]). To represent this, the natural choice is a Normal distribution. We take our representative sample of proteins from the PDB, and compute their neighbouring Cα distances. To this data, we use non-linear least squares to fit a Gaussian distribution. This fit is shown in Figure 4.16, with a mean of 3.80523548 and standard deviation (SD) of 0.02116009. By evaluating the log probability of the distances of neighbouring points in our model, we penalise structures where neighbouring Cα atoms get too close or far apart.

4.3.4 The non-neighbouring Cα distance penalty.

For the pairwise distances of non-neighbouring Cα atoms, our penalty represents the van der Waals radius. That is, the distance of the effective "hard shell" around

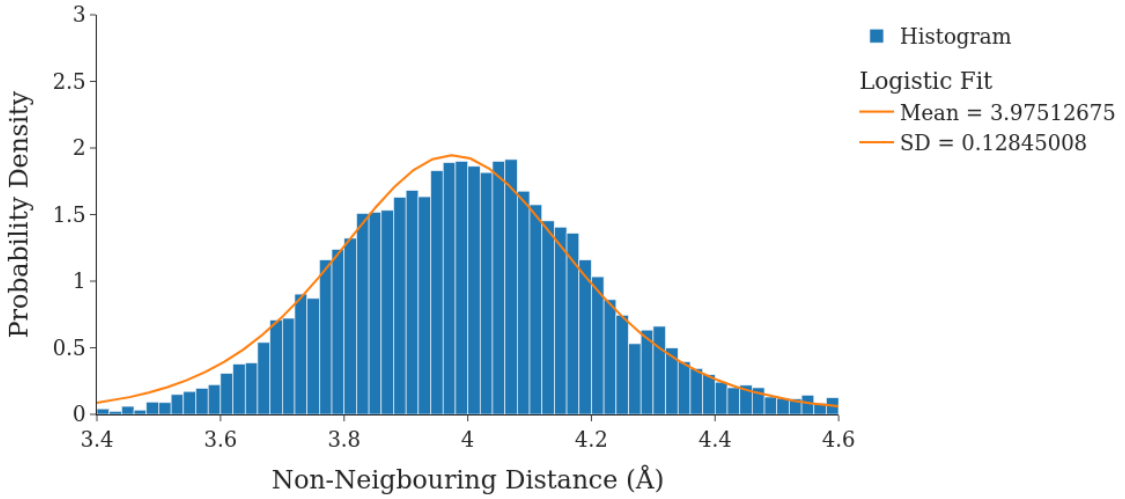
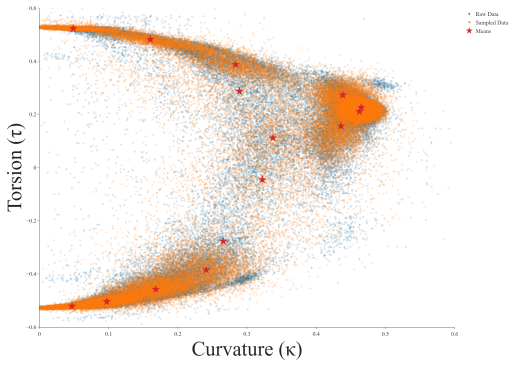


Figure 4.17: In blue, the histogram of non-neighbouring Cα distances for a representative sample of protein backbone curves. In orange, a Logistic curve fit to the histogram.

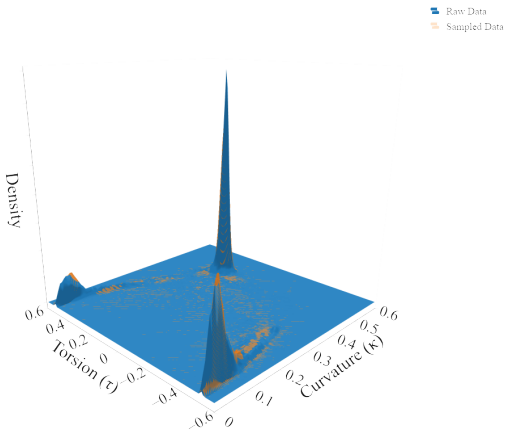
an atom which cannot be occupied by another atom. In molecular dynamics simulations, the classical choice for this is the Lennard Jones potential. Though this potential has the desired property of repulsion at small distances, it is also attractive at medium distances. In this model, we are only concerned with the points representing the Cα atoms not getting too close together, we do not want to encourage bonds through attraction at medium distances. To this end, we opt for a Logistic distribution, with mean computed from all non-neighbouring Cα distances for our sample of protein backbone curves. In Figure 4.17 we see the fit of the Logistic curve to the distribution of non-neighbouring Cα distances, its mean is 3.97512675 and standard deviation 0.12845008. Evaluating the log probability of this Logistic CDF, we penalise non-neighbouring residues becoming too close together.

4.3.5 The curvature-torsion penalty

The implementation of the curvature-torsion constraints is the biggest change between this complementary model and the original Carbonara model. Instead of finding distributions of curvature-torsion from which we can sample values to generate curves, we will instead fit a distribution to the curvature-torsion data and penalise



A Scatter plots of the curvature torsion distribution for protein backbone curves in blue, and the GMM sampled distribution in orange. The GMM means are highlighted with red stars.



B Surface plots of the curvature torsion distribution for protein backbone curves in blue, and the GMM sampled distribution in orange.

Figure 4.18: The curvature torsion distribution for a representative sample of protein backbone curves compared to the GMM distribution.

low probability changes to the backbone curve according to this distribution. To do this, we fit a Gaussian Mixture Model (GMM) with 15 components to the general curvature-torsion data, that is, without separating the data into secondary structure types. In Figure 4.18 we see the fit of this GMM to the distribution of curvature and torsion across the PDB. In Figure 4.18A we see the location of the means for each Gaussian distribution in the mixture. Again, the sharpness of this distribution is more visually apparent in the surface plot Figure 4.18B. The GMM effectively captures the peaks of the distribution corresponding to the secondary structure types, whilst allowing the variation of curvature-torsion values within linkers.

4.3.6 The global *acn* penalty

We cannot use the SKMT algorithm exactly as detailed in Chapter 3 due to the computational complexity relative to this gradient descent algorithm. In particular, performing the entanglement check in the algorithm is an $O(n^2)$ loop over all pairs of points of the curve. As a result, we must find an approximation to this algorithm and adjust the *acn* penalty accordingly. Recall that the broad approach of the SKMT algorithm is to replace the rigid SSEs by their start points, and to reduce linkers to as few points as possible to maintain non-local entanglements. An approximation

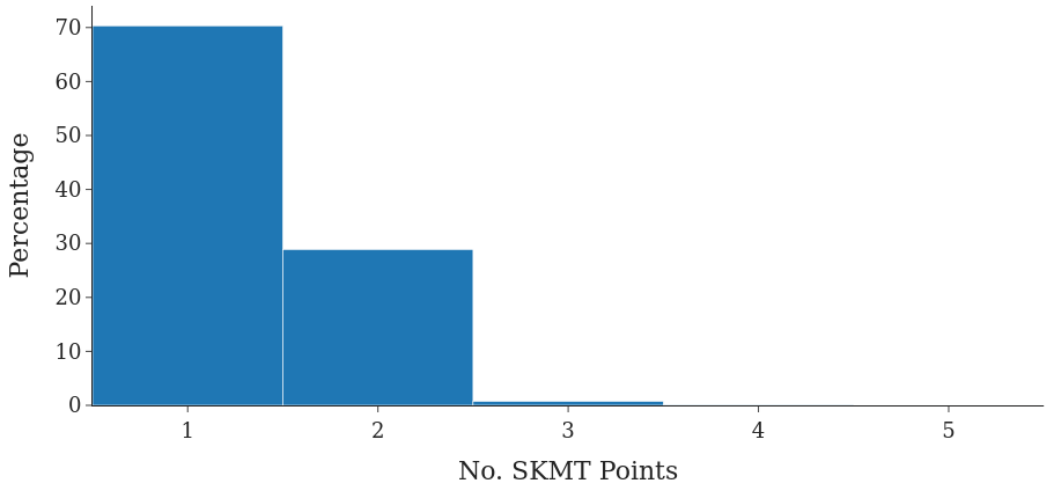


Figure 4.19: A histogram showing the distribution of the number of SKMT points needed to represent linkers.

to the SKMT curve would be a curve consisting of the start points of the α-helices and β-strands, and some to be determined number evenly spaced points from each linker.

In Figure 4.19 we see that 70% of the linkers are reduced to simply their starting point by the SKMT algorithm. However, for the remaining 30% of linkers, at least one other point is left to ensure the preservation of non-local entanglement. For a safe approximation of the SKMT algorithm in this model, we will replace each linker with its starting and middle point. For our representative sample from the PDB, we find the approximation to the SKMT smoothed backbone curve of each protein, and compute its *acn*. This distribution of *acn* is shown in Figure 4.20, with a lower bounding curve fit to these data shown in green. We also plot the original lower bounding curve from Figure 4.2 to show the difference due to this approximation. We construct the *acn* penalty as a logistic distribution fit to this lower bounding curve, with a length-dependent mean and standard deviation. Evaluating the log probability of this Logistic CDF will ensure that the curve does not open out too much to become unrealistically entangled.

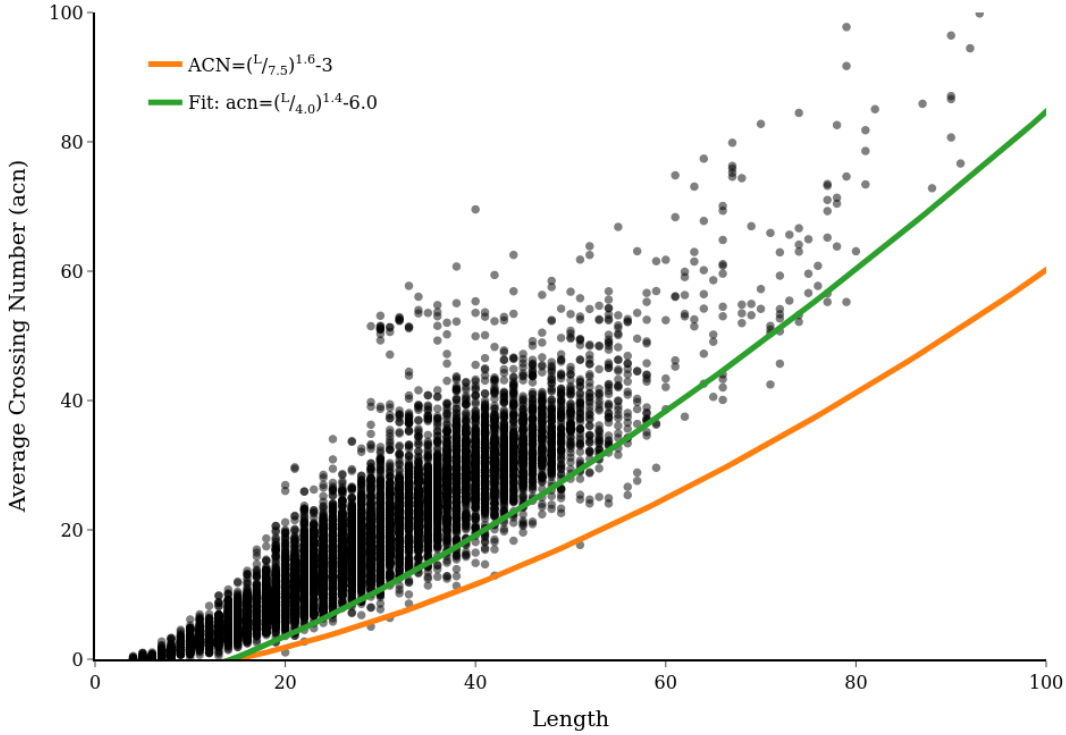


Figure 4.20: The distribution of *acn* against length for pseudo-SKMT smoothed backbone curves. In orange we plot the *acn* bound from the original SKMT data. In green we plot a new lower bounding curve fit to this data.

4.3.7 The subdomain *acn* penalty

With the tightly constrained curvature-torsion distributions, we have control over the secondary structure of the backbone. Now, with the empirical bound on the *acn*, we can also penalise unrealistic tertiary structures. However, as we have shown in Chapter 3, there exist common structural motifs between these two scales, on the super secondary level. It would therefore be useful to be able to penalise a structure whose subdomains become unrealistically entangled in a prediction. Indeed, it was shown in [69] that *acn* of protein subdomains of a given length scales similarly to that of proteins of that length. We develop this work using the SKMT backbone representation. As an example, consider the two protein backbone curves shown in Figure 4.21. In Figure 4.21A, we see the backbone of PDB entry 1CLL, which is an experimentally determined structure for Calmodulin in the presence of calcium (see Section 1.3.1 for more detail on this protein). In Figure 4.21B we see a Carbonara predicted model for this protein in solution. Despite this conformation having an *acn* above the lower bounding curve and not triggering the penalty in Carbonara, the C-terminal domain (at the top of Figure 4.21B) has come completely apart. This highlights how a protein can contain enough entanglement globally, despite having regions with little to no entanglement. It is exactly these situations for which a bound on the *acn* of subdomains will be vital for maintaining realistic entanglements on all scales.

To measure the entanglement of subdomains, we use the *acn* fingerprint. The *acn* of subsections of a given length k will be contained along the diagonals of the *acn* fingerprint matrix. For example, the principal diagonal entries are of the form $acn(\mathcal{C}_{i,i+4})$ for $i = 1, \ldots, n-4$. In order to derive a lower limit on the entanglement of subdomains, we will take the minimum of the *acn* of all subsections of lengths $k = 4, \ldots, 100$ across our sample of SKMT smoothed backbones of proteins from the PDB. To this data, we fit a lower bounding curve using non-linear least squares regression. We restrict this to structures whose full SKMT smoothed backbone has *acn* above the original lower bound curve, since this is the exact case we are looking to constrain with this new bound. That is, a predicted model whose full backbone appears realistically entangled however it has a subdomain which is unrealistically

A Cartoon representation of Calmodulin from PDB entry 1CLL

B A poor Carbonara predicted model for Calmodulin

Figure 4.21: An example of a protein backbone curve whose global *acn* may be above the bound, but contains subdomains with unrealistic entanglement.

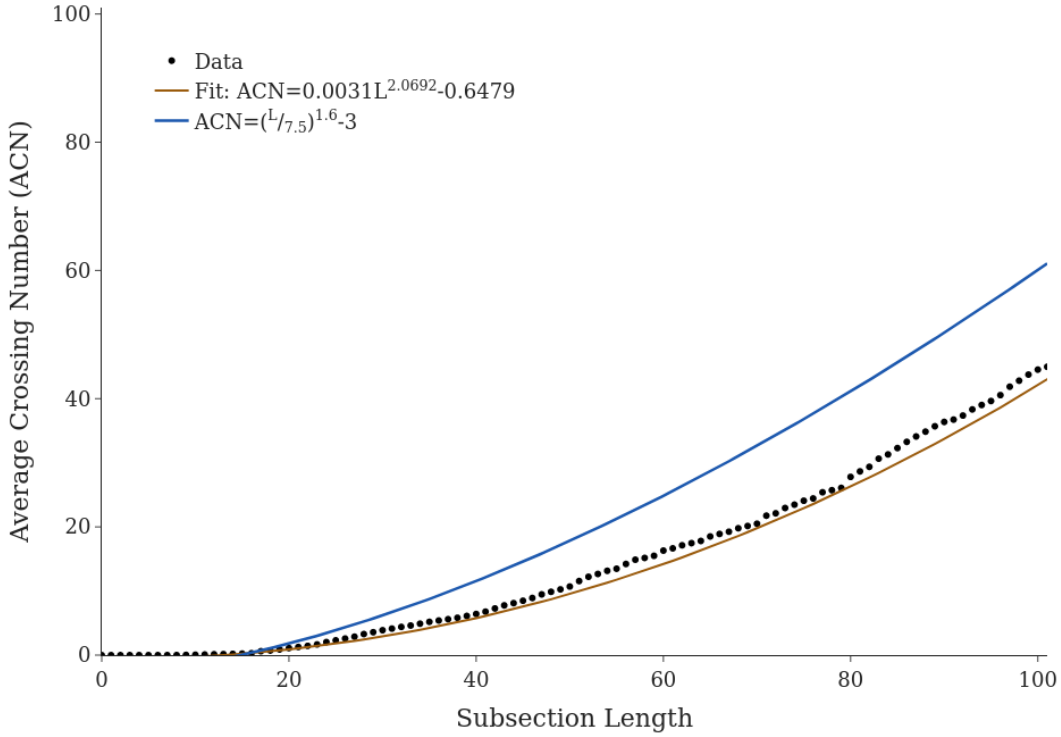
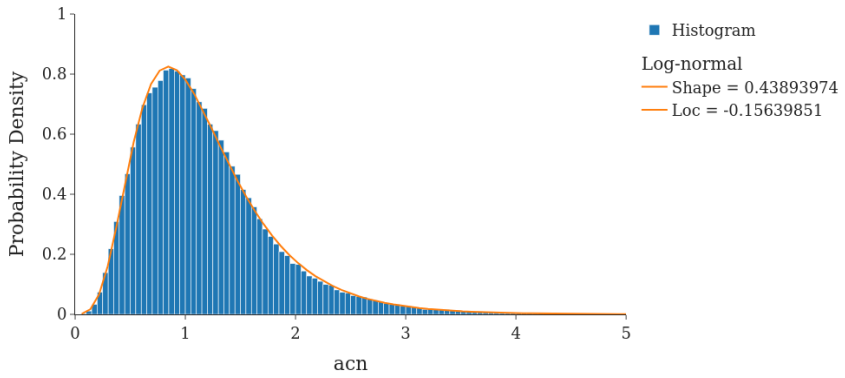


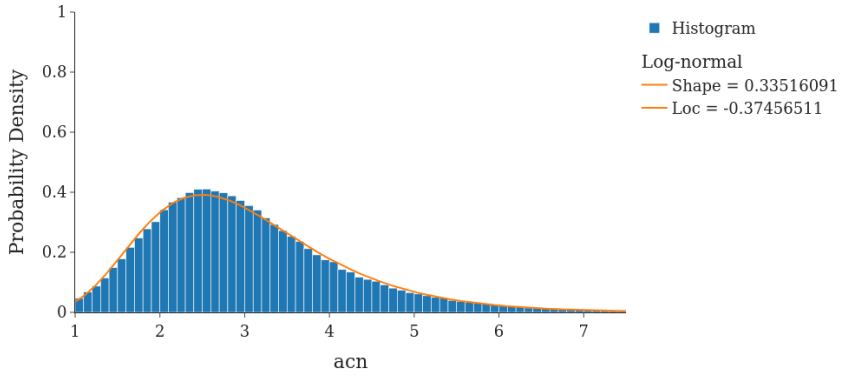
Figure 4.22: In black, the distribution of minimum *acn* against subsections lengths. In brown, a curve fit to this data using least squares. In blue, the *acn* bound for full structures from Figure 4.2.

entangled. In Figure 4.22 we see the distribution of these minimal *acn* values for each subsection length. By plotting the *acn* bound for full structures of Figure 4.2, we see that subdomains can have significantly lower *acn* than complete structures of the same size. Of note in this plot is that some subsections with length $L \leq 20$ have essentially zero *acn*. To construct a penalty for unrealistically entangled structures on this scale, we cannot simply use this lower bound.

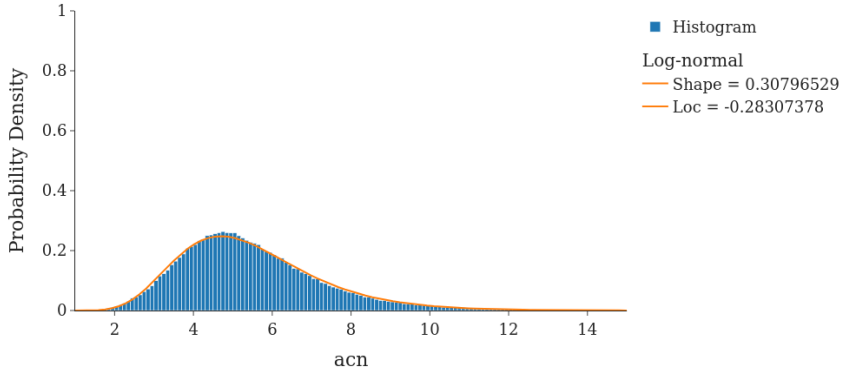
We see in Figure 4.23 that although subsections of length $L \leq 20$ can potentially have zero *acn*, the more favourable conformations are more entangled. For an effective penalty on the entanglement of subdomains, we penalise subsections whose *acn* is away from the peaks of these distributions. Since similar plots to those in Figure 4.23 were seen for lengths $L = 5, 10, \ldots, 45, 50$, we implement this penalty based on a log-normal distribution whose shape and location are functions of the length of the subsection.



A In blue, a histogram of the distribution of *acn* for all subsections of length 10. In orange, a log-normal distribution fit to this histogram.



B In blue, a histogram of the distribution of *acn* for all subsections of length 15. In orange, a log-normal distribution fit to this histogram.



C In blue, a histogram of the distribution of *acn* for all subsections of length 20. In orange, a log-normal distribution fit to this histogram.

Figure 4.23: Histograms of the *acn* of all subsections of length $L = 10, 15, 20$, with log-normal distributions fit to these histograms

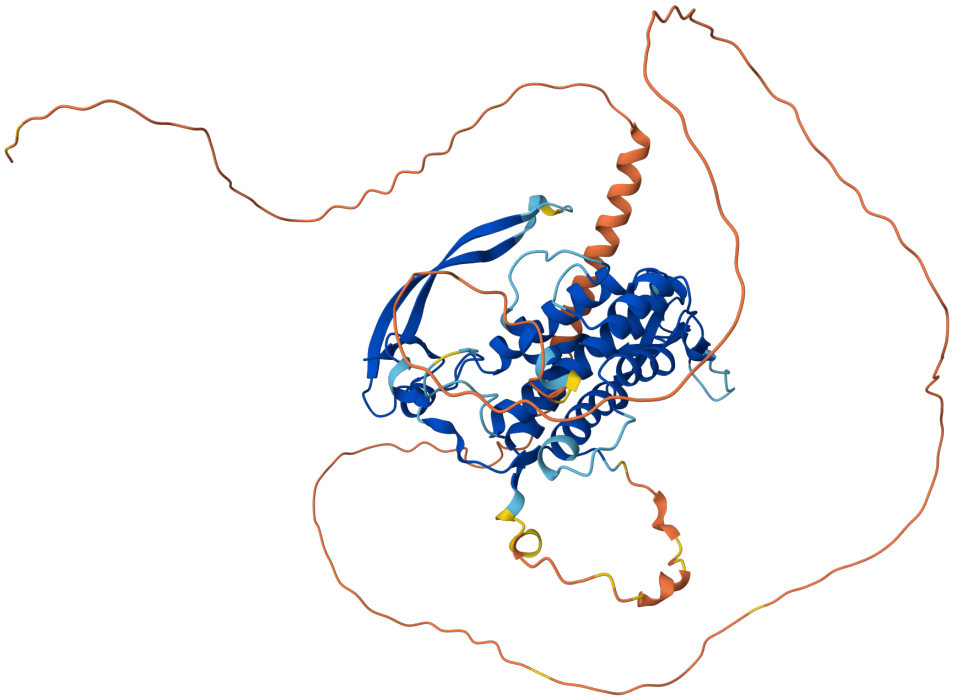


Figure 4.24: The AlphaFold predicted backbone curve for protein Wnt4 from Drosophila melanogaster (UniProt: D2NUH8). Regions of high confidence are shown in blue, and regions of low confidence are shown in orange/yellow.

4.3.8 Wnt4 - A proof of concept for this model

As an initial proof of concept for this model, consider the AlphaFold predicted model for protein Wnt4 shown in Figure 4.24. This predicted model has many regions of low confidence; in particular, the first 220 residues are essentially unordered. For residues that AlphaFold cannot accurately predict, it opts for a "safe" approach of forming a long loop which does not interact with the rest of the conformation. From the bounds on entanglement we have determined in this thesis, this conformation is evidently unrealistic. This is another example of a protein where globally the *acn* is sufficient, however it contains subsections (the poorly characterised loops) with little to no entanglement. We will therefore aim to optimise this backbone curve using our complementary model, with a specific focus on forming a more compact globular structure. This optimisation will be based on the geometrical penalties outlined above, as SAXS data is not available for this protein and a scattering

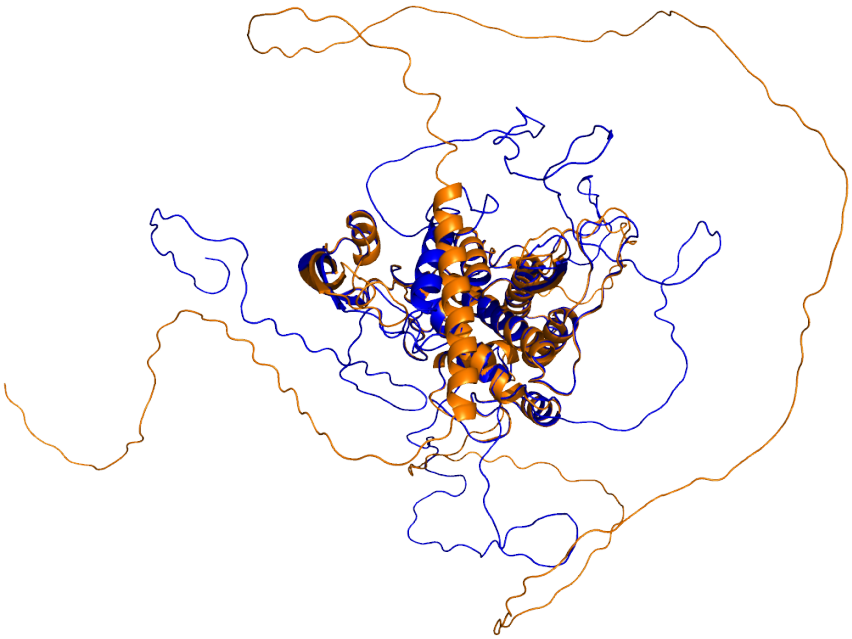
penalty is not yet implemented in this model.

In Figure 4.25 we see the AlphaFold predicted structure for Wnt4 aligned with an optimised backbone curve from our model. It is clear from visual inspection that this output structure is significantly more globular. In particular, the long looped sections which are characteristic of low confidence AlphaFold predictions have begun to form more local structure. This has also been possible whilst preserving the structural integrity of the high confidence region at the core of the protein. Using the MolProbity web server [101] to compare the model quality for these predictions, we find that our optimised model has a higher percentage of favoured rotamers (90.1% to 88%), favoured Ramachandran angles (84.4% to 80.5%) and fewer Cα geometry outliers (5.42% to 11.4%). Using the MolProbity web server [101] to compare the model quality for these predictions, we find that our optimised model has a higher percentage of favoured rotamers (90.1% to 88%), favoured Ramachandran angles (84.4% to 80.5%) and fewer Cα geometry outliers (5.42% to 11.4%). Although this is evidently still a long way from a complete prediction of the native state structure of this protein, we feel that it serves as an indicator of the potential of this model.

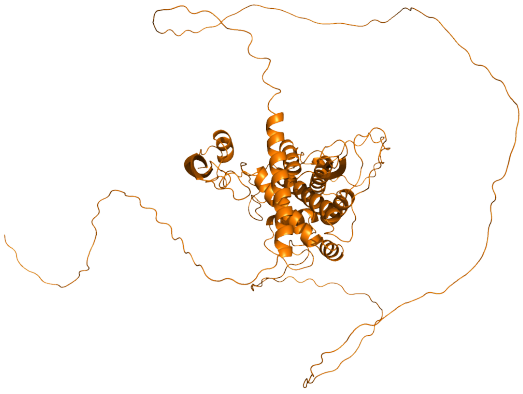
4.4 Discussion

In this chapter, we introduced the Carbonara software package, which is a development of the constrained backbone algorithm into a rapid method for refining protein structure predictions based on BioSAXS data. The key result of this chapter in relation to this development is a method for constraining the possible state space using an empirically determined bound on the entanglement of the backbone curve. By considering the *acn* of SKMT smoothed backbone curves, we found a clear lower bound on the expected amount of entanglement of the backbone curve relative to its secondary structure. We implemented this lower bound as a penalty during fitting in the Carbonara routine, and using Human SMARCAL1 as an example showed that this produces biologically plausible conformations.

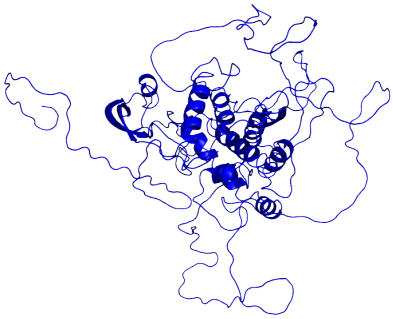
We presented an example Carbonara use case that addresses the changing landscape of tertiary structure prediction. With AlphaFold trained on crystallographic



A In orange, the Cα backbone curve of the AlphaFold predicted structure of Wnt4. In blue, the Cα backbone optimised by our model.



B The Cα backbone curve of the AlphaFold predicted structure of Wnt4.



C The optimised Cα backbone curve of Wnt4.

Figure 4.25: The AlphaFold predicted structure for Wnt4 against the optimised backbone from our model.

data, its predictions may be a poor fit to the experimental data of a solution scattering experiment where a protein is dynamic or adopts multiple conformations. We presented domain IV of DnaA from Bacillus subtilis as a guiding example, with Carbonara predicted models providing an improved fit to the scattering data while also highlighting the main flexible region. Further investigations via Molecular Dynamics simulations would be required to make a confident prediction that this new conformation is stable and biologically plausible. This work can be done as part of the wider Carbonara pipeline; however, it is outside the focus of this thesis.

We then discussed the limitations of the Carbonara model in relation to its Monte Carlo approach. By sampling new geometries for linker subsections, the model can rapidly produce large scale conformational changes which would be inaccessible by Molecular Dynamics simulations. However, when subtler changes to the backbone are required, the model can have difficulty finding an improved conformation. To this end, we presented the framework for a complementary model to Carbonara which takes a gradient descent approach. We have shown that this model can produce a more realistically entangled backbone curve, taking a low confidence AlphaFold prediction as input and moving towards a more globular conformation. The results of this optimisation are not to be considered a complete prediction of the structure for this protein, and are included purely as an example of the geometrical constraints producing sensible changes to an unrealistic backbone.

The early results of this model are promising; however, it is evidently still in the early stages of development. The most immediate objective is to implement an effective scattering penalty for this model. This work is currently in progress but is not yet reportable. With this in place, the addition of context driven constraints which mirror those in Carbonara, such as pairwise distance constraints and fixed sections, would allow for the inclusion of this model into the full prediction pipeline. This model could then provide a bridge between the original Carbonara model and the computationally expensive Molecular Dynamics simulations.

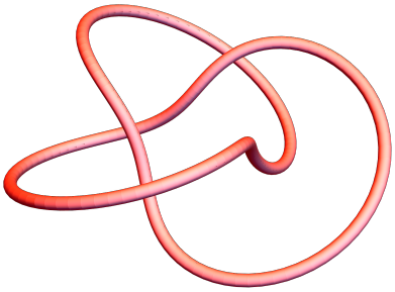
CHAPTER 5

A writhe based similarity metric for flexible structure comparison

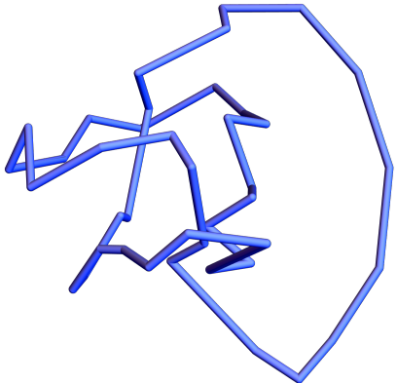
In this chapter we will define a writhe based metric for SKMT smoothed backbone curves. This writhe based metric is motivated by the flexible nature of proteins in solution. We will show through an example the types of similarity we are looking to quantify with this metric and where they are important. In particular, the aims for the metric are the following:

- The metric should classify two structures as similar if they can be distorted into each other without significantly changing the nature of the tertiary fold.
- The metric should highlight specific arrangements of the secondary structure elements, say, large-scale helical super secondary structures.
- The metric should also detect potentially meaningful structural similarities that are missed by other standard measures.

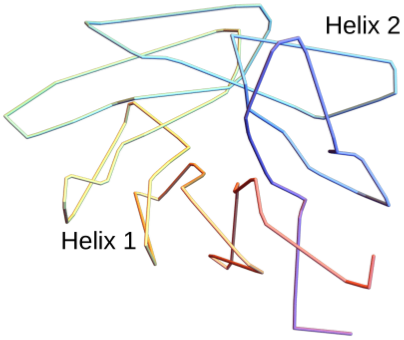
This chapter will contain some elements of shared work with Christopher Prior that have appeared in [3].



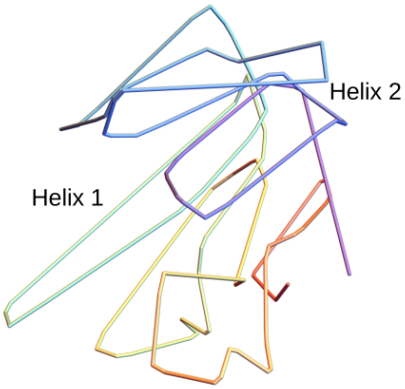
A A smooth trefoil curve.



B A distorted trefoil curve.



C The smoothed Cα backbone of the monomer unit of 2-deoxyribose-5-phosphate aldolase (PDB: 1P1X)



D The smoothed Cα backbone of an oxidoreductase, yciK from *E.coli* (PDB: 3F1L)

Figure 5.1: Depictions of the notion of structural similarity we seek to quantify in this study. In both examples the topological similarity would be missed by distanced based metrics.

5.1 The need for flexible similarity metrics

As we have seen in the previous chapter, proteins in solution can adapt multiple conformations and exhibit a high degree of flexibility. As a result, the resolution of the information that we commonly obtain from scattering experiments is not sufficient to understand the exact position of each amino acid. As a result, any notion of similarity between output structures should be robust to these small scale variations.

An example of the types of similarity that we are looking to characterise is shown in Figure 5.1A-B. The trefoil curve Figure 5.1A is knotted in the sense that it must be cut to deform it into a circle. The curve shown in Figure 5.1B can be obtained by locally distorting Figure 5.1A continuously without the curve crossing itself,

that is, without any such cutting. Thus in some sense they are folded in a similar fashion, but this would not be captured by the standard distance-based metrics used for rigid structure comparison. A more pertinent example with real protein structures is shown in Figure 5.1C-D where the SKMT smoothed Cα backbones of two proteins which have different CATH classifications, a TIM barrel (Figure 5.1C) and a Rossmann fold (Figure 5.1D), can be seen to bear a striking similarity up to such a distortion. They have clear helically coiled domains with the same number of helical loops in each domain and similar relative orientations of these domains.

This specific example is more than just a curiosity. In fact, the results of [28] indicate a very close link between these two CATH domains on the primary sequence, or even the evolutionary level. In [28] the authors used directed evolution techniques in attempts to design a TIM barrel structure, but instead found that a Rossmann fold-like structure was produced. When these results were tested against the contemporary state-of-the-art computational techniques, none were able to predict the produced structure, and most agreed that the sequence should indeed produce a TIM barrel conformation. Both TIM barrels and Rossmann folds have a β sandwich structure formed from anti parallel β-strands whose cross-bonds provide stability to the structure. The helical subsections of these two structures result from this β sandwich motif. We have already seen that a large-scale helical conformation is highly prevalent in a wide variety of protein structures, often of a similar scale across a variety of CATH families. We believe our tools for identifying similar size helical structures could provide insight into de-novo design methods.

5.2 Writhe based similarity metric

5.2.1 Definition

The visual similarity between these globally helical structures opens the question of whether we can identify other surprising tertiary structural similarities by comparing writhe profiles. To this end, we propose a writhe-based similarity metric for subsections $\mathcal{C}^1_{ij}$ and $\mathcal{C}^2_{kl}$ of the SKMT smoothed curves $\mathcal{C}^1, \mathcal{C}^2$. We fix $l - k = j - i$ so that we compare subsections of the same size (even if the whole SKMT curves

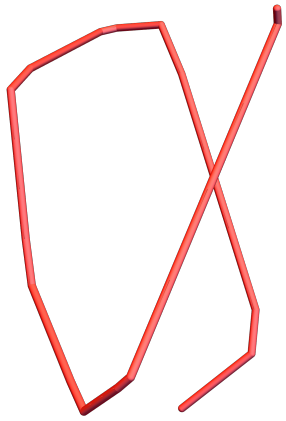
themselves may be different in size). We measure the similarity of $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl})$ as follows:

$$S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl}) = \frac{1}{j-i-4} \sum_{m=4}^{j-i} \frac{1}{0.24m} |Wr(\mathcal{C}^1_{i,i+m}) - Wr(\mathcal{C}^2_{l,l+m})|. \tag{5.1}$$

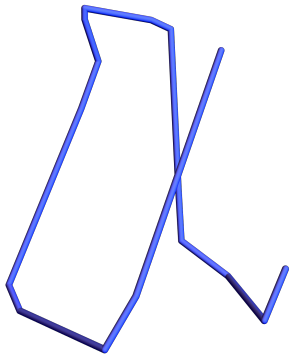
This metric measures the mean absolute difference in the writhe of all subsections $\mathcal{C}^1_{i,i+m}$, $\mathcal{C}^2_{l,l+m}$ relative to the typical linear writhe growth $0.12L$ we have seen for the SKMT smoothed protein backbones in Chapter 3. The factor features as $1/(0.24L)$ to account for the maximal observed difference of subsections of opposing sign writhe as a function of their length. We only consider $m \geq 10$ because the value of writhe for subsections smaller than this is not meaningful. This metric is applied to all sub-sections of similar size of the two curves, with a minimum length of the sub-section of 10 to focus on relatively large scale similarities. We then select the largest disjoint subsets of the two molecules which have $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl})$ less than some specified tolerance s_0. For example, a value of $s_0 = 0.05$ would indicate that the average difference is less than 5% of the typical empirically observed growth in writhe difference. For context, a visual depiction of the level of similarity we expect for $s_0 = 0.05$ and $s_0 = 0.1$ for some example subsections is shown in Figure 5.2.

The value s_0 for which sections that satisfy $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl}) < s_0$ are classified as similar is an important one. In Figure 5.3 we chart the percentage similarity as a function of s_0 for the Rossmann Fold 3F1L and the TIM Barrel 1P1X up to $s_0 = 0.2$. There is a sharp increase between 0.02 and 0.05, then a more steady increase. By 0.1 all of 3F1L (the smaller molecule) is considered the same as about 75% of 1P1X.

We will now showcase some example uses of this metric for structural comparison, before moving onto performing a clustering using a full pairwise sweep across our database. These examples are included in an interactive iPython notebook at https://github.com/arronelab/SWRITHE to allow researchers to benefit from these tools.

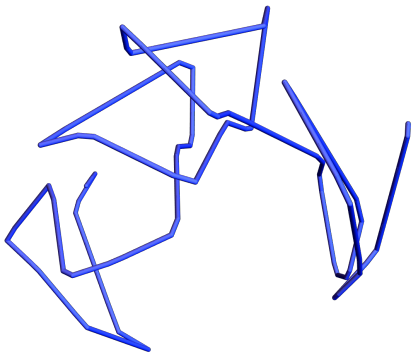


A A section of curve $\mathcal{C}^1_{ij}$ for which $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl}) = 0.05$ when compared to the curve shown in panel (b).



B A section of curve $\mathcal{C}^2_{kl}$ for which $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl}) = 0.05$ when compared to the curve shown in panel (a)

C A section of curve $\mathcal{C}^1_{ij}$ for which $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl}) = 0.1$ when compared to the curve shown in panel (b).

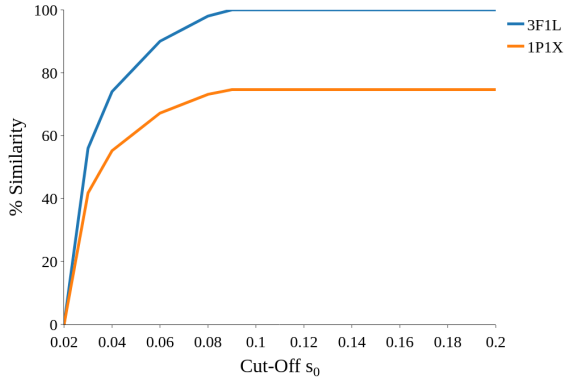


D A section of curve $\mathcal{C}^2_{kl}$ for which $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl}) = 0.1$ when compared to the curve shown in panel (b).

Figure 5.2: Examples of curve sections sharing a 5% similarity by our comparison metric in (a) and (b) and a 10% similarity in (c) and (d). Note the single helical loop in (a) and (b) are very uniform, whereas the four helical loops in (c) and (d) are less coherent, especially for (d).

A The SKMT smoothed backbone of 1P1X, with the mutually similar section with 3F1L highlighted in green.

B The SKMT smoothed backbone of 3F1L, with the mutually similar section with 1P1X highlighted in green.



C The percentage similarity of 1P1X and 3F1L as a function of the cut-off parameter s_0

Figure 5.3: Visualisations of the similarity metric $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{kl})$ for the example Rossmann Fold (3F1L) and TIM Barrel (1P1X) domains.

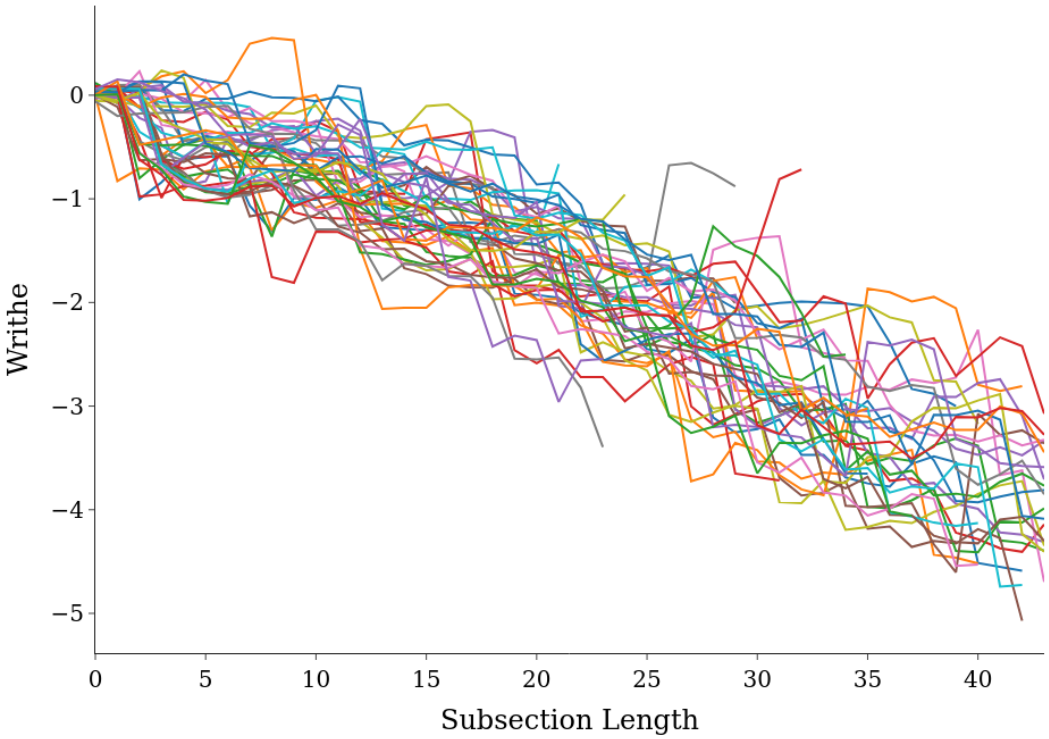


Figure 5.4: The writhe profile of the matched subsections for the cluster of proteins with $L \in [40, 66]$ and super-linear *acn*.

5.2.2 The super-linear *acn* cluster

In Chapter 4 we saw the distribution of *acn* for a representative sample of SKMT smoothed protein backbones. In this distribution, there was a cluster of proteins whose *acn* had super-linear growth. To investigate this cluster further, we compute our similarity metric for each of the proteins in this sample. We find that the mean pairwise similarity for this cluster is 0.83 indicating a high degree of similarity.

In Figure 5.4 we plot the writhe profiles for the matched subsections for this subset of proteins. The metric identifies a shared linear writhe profile amongst this sample, indicating a highly ordered helical global structure. Indeed, 62% of these proteins have a β-barrel architecture according to CATH, compared to just 3.4% of the total sample having this architecture. Further to this, of the 23 proteins in our dataset whose CATH topology is "Green Fluorescent Protein", 20 of them belong to this super-linear cluster. For a full review of these proteins see [102], of particular

note here though is that their highly ordered β-barrel structure is essential to their thermal and chemical stability.

5.2.3 The Rossmann fold and TIM barrel relationship

To investigate how consistent the apparent relationship between Rossmann fold and TIM barrel domains is, we applied this metric to compare our example Rossmann Fold domain 3F1L to all other proteins in our data set. We set $s_0 = 0.05$ and restrict ourselves to the cases where both structures are classified as similar for $> 80\%$ of their length, so that we are looking at structures that are globally very similar to 3F1L. We find 112 such cases. A comparison with the CATH classification of these proteins shows that 62.5% of them are classified as Rossmann folds (as expected) and 6.3% as TIM barrel domains, seemingly strengthening the structural similarity relationship between these fold types. For context 18.3% of the full database are classed as a Rossmann fold domain, and 3.4% as a TIM barrel.

In Figure 5.5 we plot the matched subsections of the writhe profiles of this set of proteins. As expected, we see that a shared linear growth in writhe is identified, corresponding to the globally helical arrangement of the secondary structures which is typical of these CATH topologies. We also note here that there are subsections of 3F1L's writhe profile which contain the roadie wrap motif described in Chapter 3, for example from lengths $L = 30 - 40$. This will be discussed further when we perform clustering on our dataset using this metric in Section 5.3

5.2.4 Identifying nearly knotted proteins

We also performed a similarity sweep across the database for the trefoil knotted protein 2RH3 highlighted in Figure 3.5C. This produced just one match above 80% similarity, PDB entry 7YTT. The matched subsections $(\mathcal{C}_{1,19})^1, (\mathcal{C}_{1,19})^2$ cover 82% and 90% of 2RH3 and 7YTT respectively. The matched subsections of the writhe fingerprints of these two proteins are shown in Figure 5.6. One can see that these profiles are very similar, except for the sharp (negative) jump in writhe for 2RH3 with the inclusion of its final subsections. This sharp jump is due to the knotted

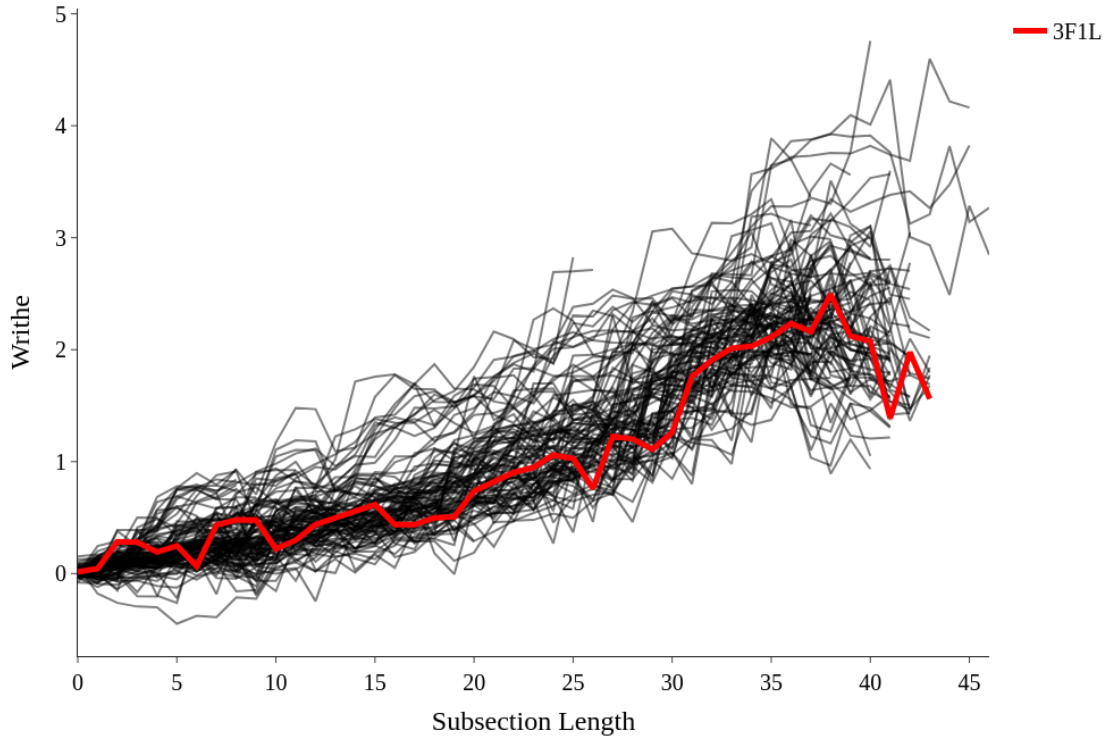
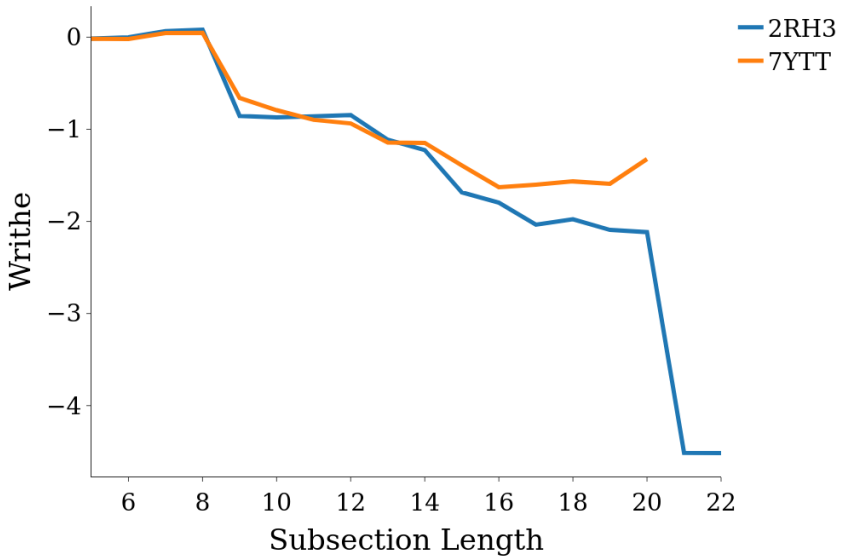
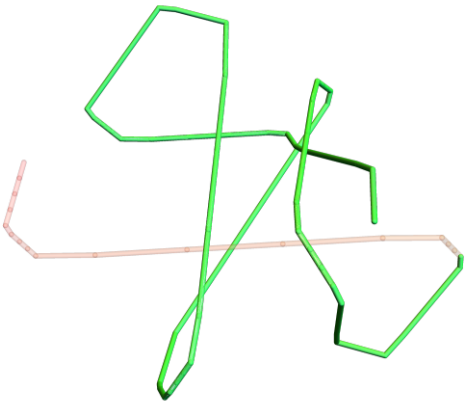


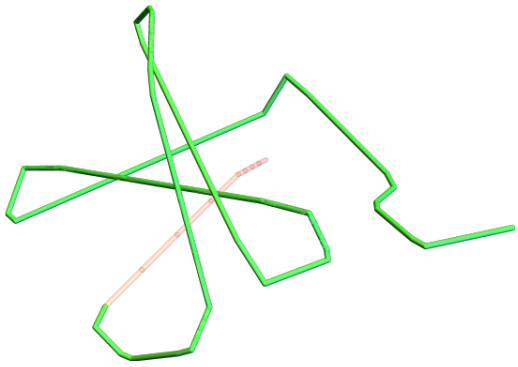
Figure 5.5: The writhe profiles of the matched subsections for proteins globally similar to PDB entry 3F1L.



A The matched subsections of the writhe fingerprints for the trefoil knotted 2RH3 and its single similarity match from our database, 7YTT.



B The smoothed backbone of 2RH3, with the mutually similar sections to 7YTT highlighted in green.



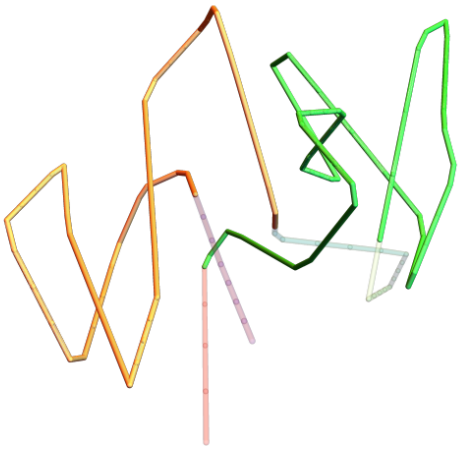
C The smoothed backbone of 7YTT, with the mutually similar sections to 2RH3 highlighted in green.

Figure 5.6: The matched subsections of the writhe fingerprints and mutually similar sections of the trefoil knotted 2RH3 and unknotted 7YTT.

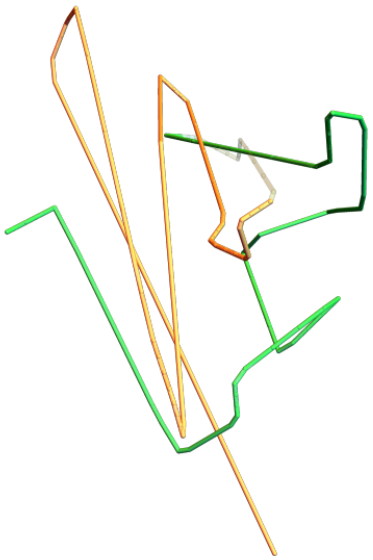
nature of this backbone, with the C-terminus threading through the rest of the structure to complete the trefoil. By contrast we find, using the KnotProt identification software [47], that the backbone of 7YTT is unknotted. In Figure 5.6B we see for 2RH3 the C-terminus (shown in translucent red) threads through the green section of curve (which is mutually similar for 2RH3 and 7YTT). In Figure 5.6C however, the C-terminus of the 7YTT backbone resides on the outside of the structure, which prevents it from being classed as a knotted structure. The metric has therefore identified two structures that differ only by their C-terminus threading. This difference would be missed by classical distance based metrics, as explained in Chapter 2. Similarly, the CATH classification misses the apparent similarity; in fact, 7YTT has no classification. This highlights the potential in this comparative metric for identifying similar protein folds which are missed by other classification methodologies.

5.2.5 A lonely Rossmann fold

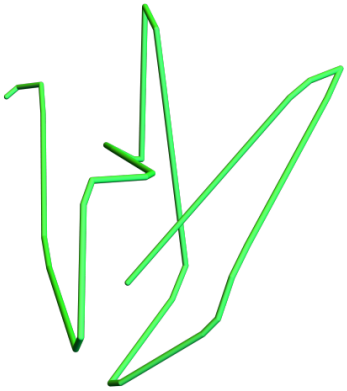
To conclude this section on the similarity metric, we consider an example Rossmann fold domain which was not classified as similar with 3F1L above, namely PDB entry 4QFB. Performing a sweep across our database for structures with over 80% similarity we find just one example, PDB entry 6GN5. The matched subsections are $\mathcal{C}^1_{21,36}$, $\mathcal{C}^2_{1,16}$ and $\mathcal{C}^1_{3,16}$, $\mathcal{C}^2_{21,34}$ that cover 81% and 88% of 4QFB and 6GN5, respectively. From the highlighted sections in Figure 5.7 we can see shared large scale helical subsections, much more uniform in the case of 4QFB. Using the Search by Sequence function on the CATH website, there are no domain matches for the FASTA sequence of 6GN5. That is because, at the secondary structure level, it does not meet the strict criteria to be classified as a Rossmann fold domain. However, on the tertiary structure level, there are some clear similarities between itself and an example Rossmann fold domain.This highlights again that our comparison metric can be used to uncover potential relationships which might be missed by standard methods.



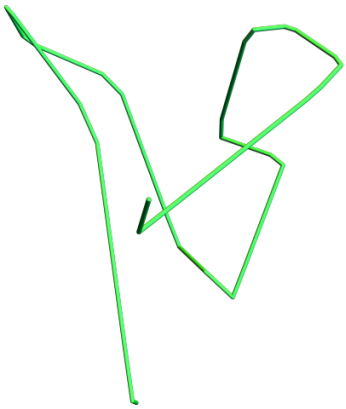
A The smoothed backbone of 4QFB, with the mutually similar sections to 6GN5 highlighted.



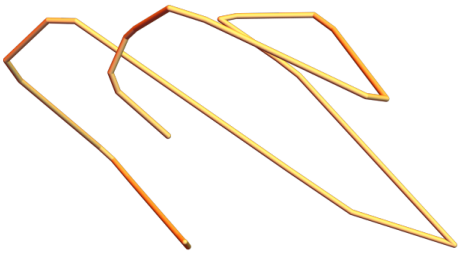
B The smoothed backbone of 6GN5, with the mutually similar sections to 4QFB highlighted.



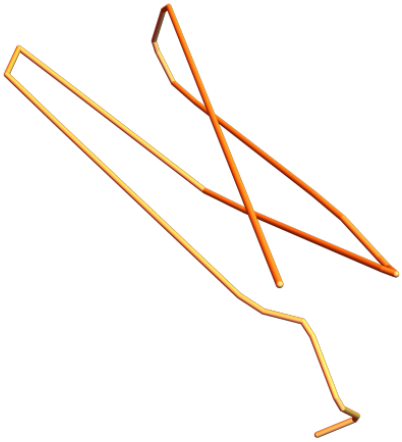
C The first matched subsection of 4QFB.



D The first matched subsection of 6GN5.



E The second matched subsection of 4QFB.



F The second matched subsection of 6GN5.

Figure 5.7: The mutually similar sections of the Rossmann Fold domain 4QFB and unclassified 6GN5.

5.3 Using the similarity metric for clustering

As we have seen in the previous sections, the writhe similarity metric can highlight shared helical geometries across various CATH families and identify protein backbones which are "nearly" knotted. To conclude this chapter, we perform some exploratory analysis on a complete pairwise sweep of our data set using this metric.

5.3.1 Constructing a pairwise similarity matrix

For every pair of SKMT smoothed backbone curves $(\mathcal{C}^n, \mathcal{C}^n)$, we find the largest disjoint subsections of the two curves with $S(\mathcal{C}^n_{ij}, \mathcal{C}^m_{kl}) < 0.1$. The percentage of the curve $\mathcal{C}^n$ (respectively, $\mathcal{C}^m$) covered by these subsections will be the entry in position (n,m) (respectively, (m,n)) of the initial pairwise similarity matrix. The resulting similarity matrix is of size 10736 x 10736, and most significantly, is asymmetric. For example, as we have seen Figure 5.3 at a similarity cut off of 0.1, the matched subsections of the Rossmann Fold 3F1L and the TIM Barrel 1P1X cover 100% of the smaller molecule 3F1L and 75% of 1P1X. Since the size and asymmetric nature of this matrix makes clustering with standard techniques not possible, we will look to filter and manipulate it to uncover clusters.

5.3.2 Returning to our motivating example

To reduce the scale of the matrix, we can filter it for proteins of length within a given range. For example, to further investigate the link between Rossmann Fold and TIM Barrel domains we can filter for proteins of a similar length to our examples from Figure 5.1, namely $50 \leq L \leq 70$. By looking at proteins of a similar length, we remove situations where a small molecule can be very similar to a small percentage of a much larger molecule. As a result, we make the similarity matrix symmetric by taking the mean of the two percentages. We then perform clustering using Hierarchical Density Based clustering using HDBSCAN [103] from Python's scikit-learn module. Given we do not know a priori how many clusters we expect, HDBSCAN is a natural choice for this problem over other methods such as K-means. HDBSCAN is an extension of DBSCAN, which clusters points according to their

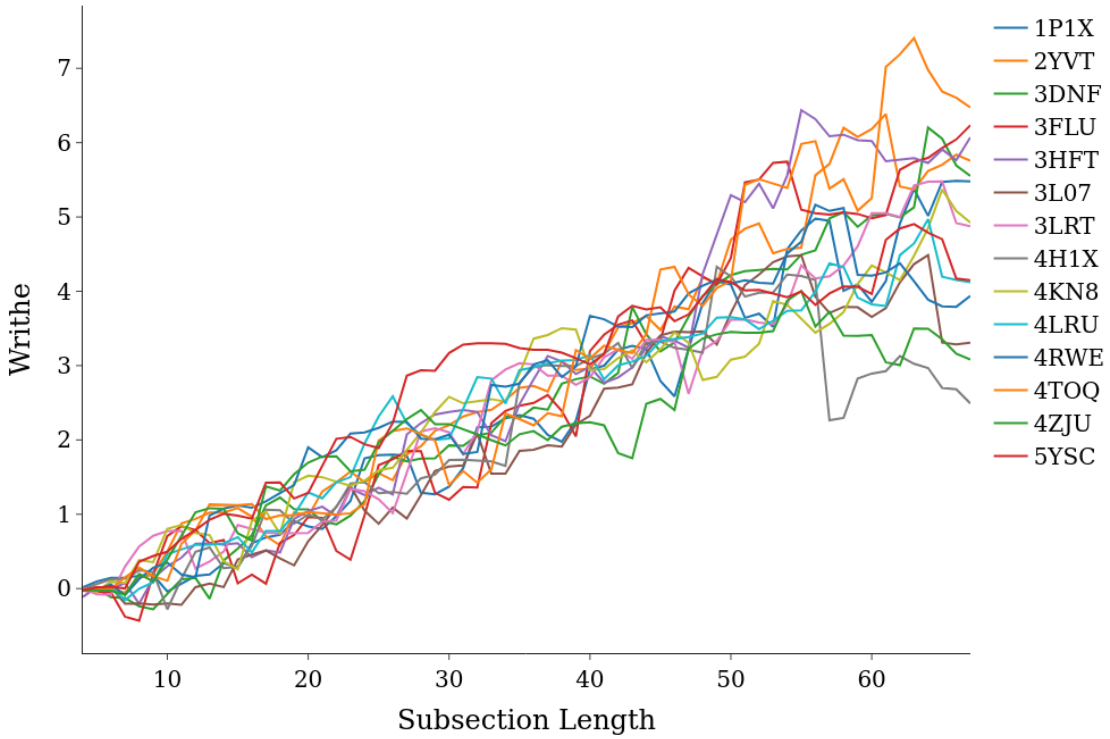


Figure 5.8: The writhe profiles for the proteins clustered with 1P1X, an example TIM Barrel.

density relative to a maximum distance threshold paramter ϵ. In HDBSCAN, a hierarchy of such clusters are constructured for various ϵ values, and then returns the most stable clusters over ϵ.

In Figures 5.8 and 5.9 we plot the writhe profiles for proteins in the clusters containing 1P1X and 3F1L. Of note initially is that despite the similarity between these two, they belong to distinct clusters when considered as part of this larger search. The difference in the uniformity of the helical coiling as seen in Figures 5.1C and 5.1D is sufficient for these to belong to different clusters. This can be seen in the writhe profiles for proteins within 1P1X's cluster being much more uniformly linear than those in 3F1L's cluster. This deviation in the uniformity of the global helical structure is not however a general fact for Rossmann folds and TIM barrels, indeed these two CATH topologies are present in both clusters.

Members of 3F1L's cluster can broadly be put into two categories. The first and most obvious are those that are matched because of their globally helical conforma-

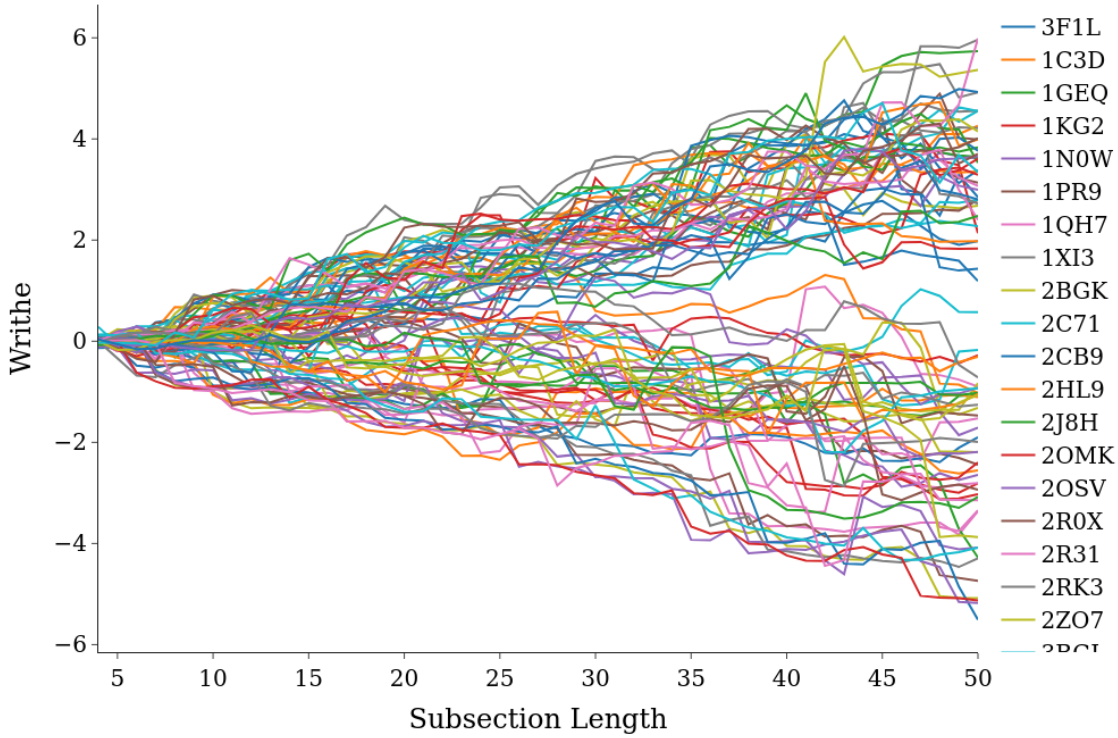


Figure 5.9: The writhe profiles for the proteins clustered with 3F1L, an example Rossmann Fold.

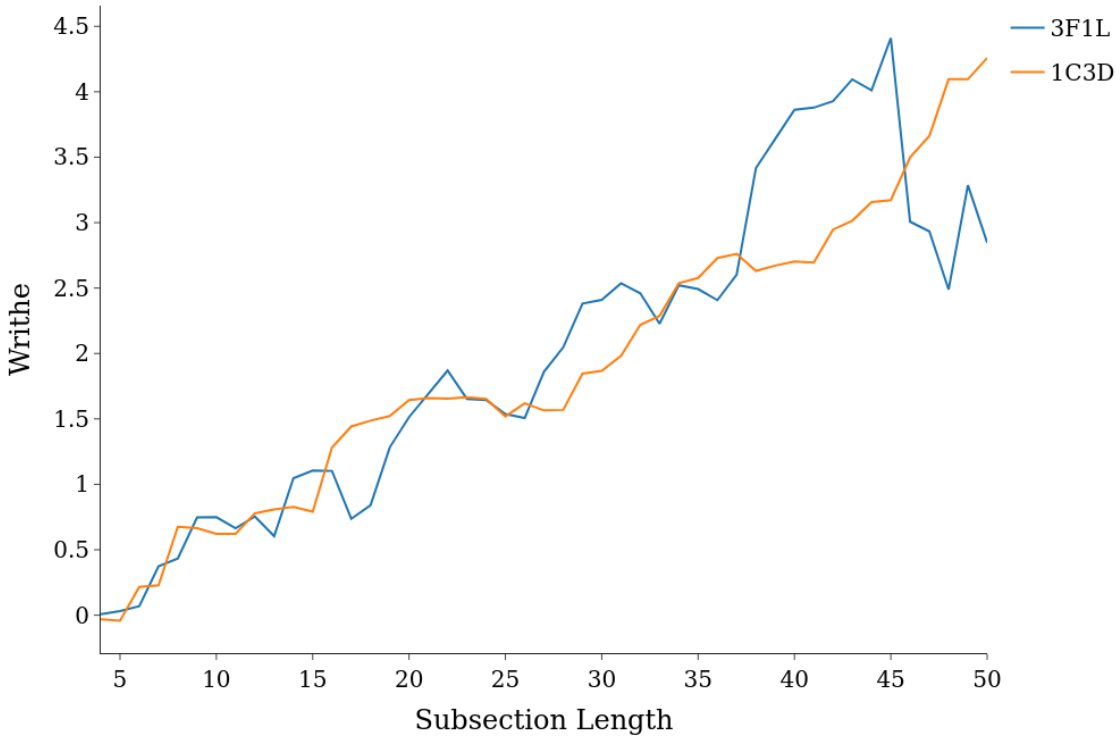
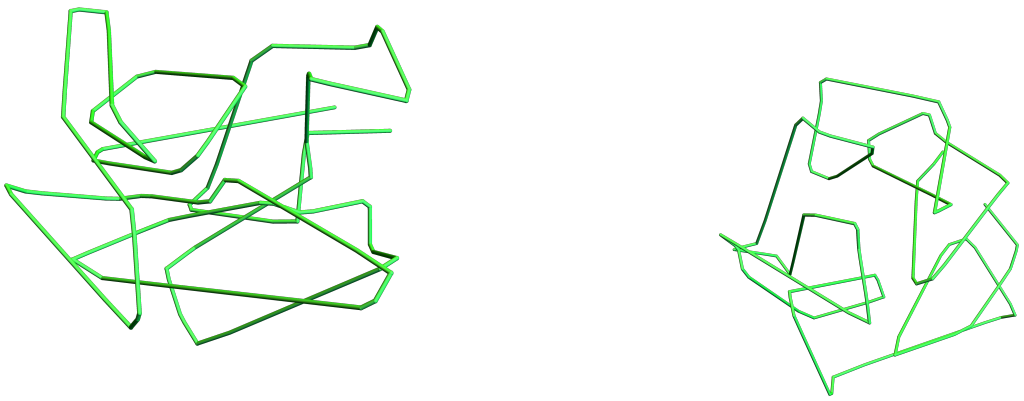


Figure 5.10: The matched subsections of the writhe fingerprints for PDB entries 3F1L and 1C3D.



A The SKMT smoothed backbone of 3F1L, with the entirety of the curve mutually similar to 1C3D.

B The SKMT smoothed backbone of 1C3D, with the entirety of the curve mutually similar to 3F1L.

Figure 5.11: The mutually similar SKMT backbone curves of PDB entries 3F1L and 1C3D.

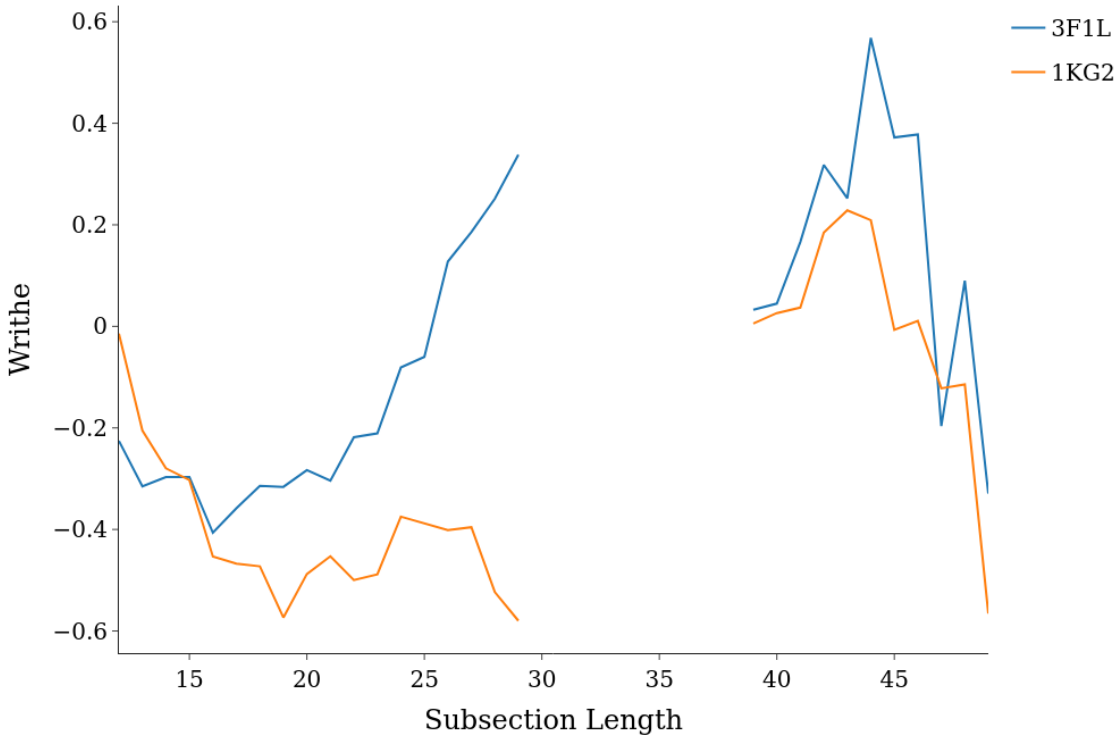
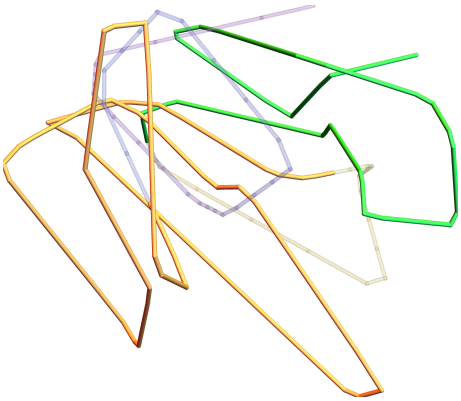


Figure 5.12: The matched subsections of the writhe fingerprints for PDB entries 3F1L and 1KG2.

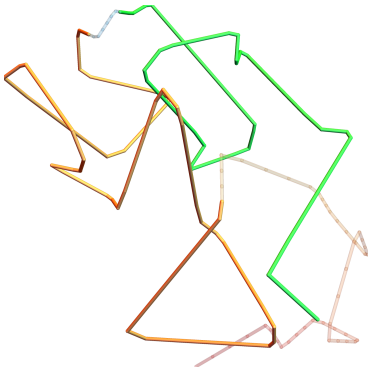
tions. An example of this type of similarity is shown in Figures 5.10 and 5.11, with the entire globally helical backbone of 1C3D being similar to that of 3F1L.

The similarities present between other members of 3F1L's cluster are a little more subtle, and lead to the lack of visual similarity between the writhe profiles of the whole molecules (not just the matched subsections) for this cluster. It is important to note that the writhe similarity metric is defined across the whole writhe fingerprint and can therefore identify similar subsections that may not be visible in the entire writhe profile. For example, consider PDB entry 1KG2 which is a member of 3F1L's cluster with a similarity score of 0.74. In Figure 5.12 we see the matched subsections of the writhe fingerprints for this pair. We have aligned these matched subsections according to their position in 3F1L's backbone.

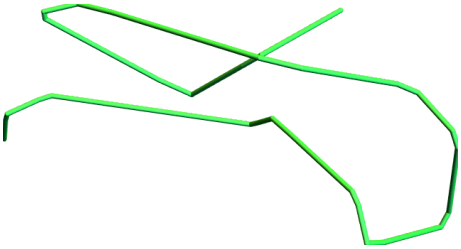
By highlighting these matched subsections in the SKMT smoothed backbone curves in Figure 5.13, we can see the similarities that this metric has identified. In particular, the first matched subsection in green is a clear shared single helical



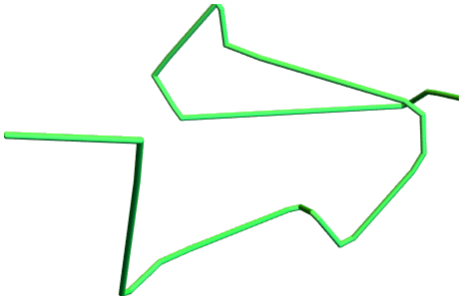
A The SKMT smoothed backbone of 3F1L, with the mutually similar sections to 1KG2 highlighted.



B The SKMT smoothed backbone of 1KG2, with the mutually similar sections to 3F1L highlighted.

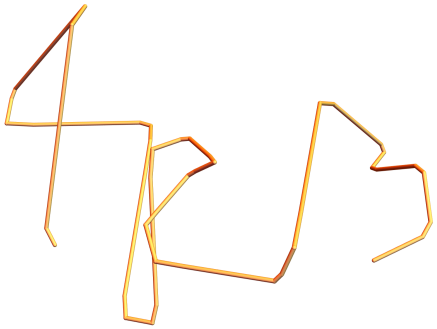


C The first matched subsection of 3F1L.



D The first matched subsection of 1KG2.

E The second matched subsection of 3F1L.



F The second matched subsection of 1KG2.

Figure 5.13: The mutually similar sections of PDB entries 3F1L and 1KG2.

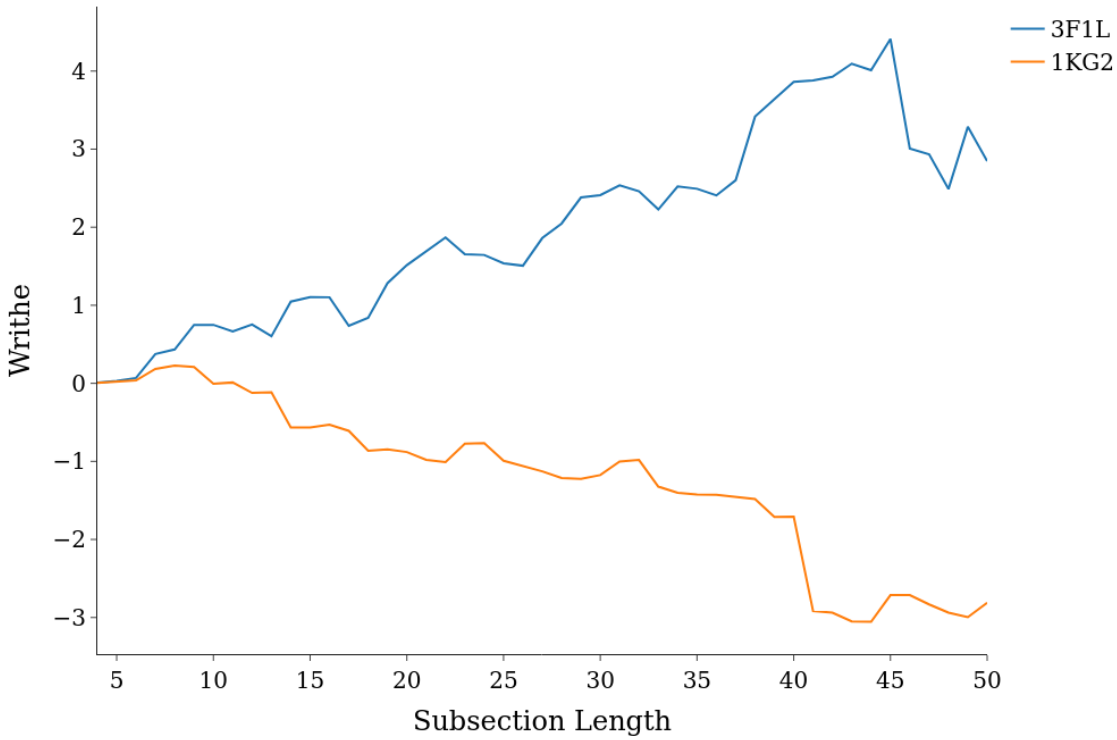


Figure 5.14: The writhe profiles of the whole SKMT backbone curves of 3F1L and 1KG2

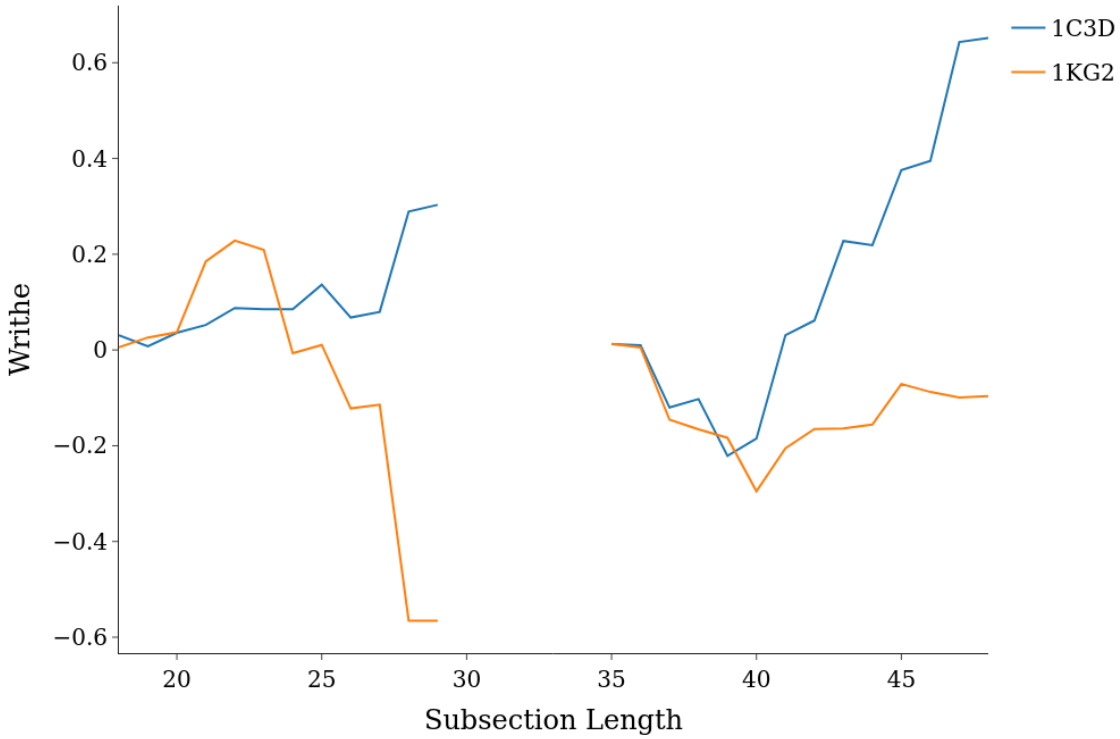


Figure 5.15: The matched subsections of the writhe fingerprints of PDB entries 1C3D and 1KG2.

loop, as part of the C-terminus for 3F1L and the N-terminus for 1KG2. The second matched subsections are much less uniform helical geometries, accounting for the middle 40% of each protein. In the case of the 3F1L, this subsection is the region connecting the two halves of the Rossmann Fold, with this folding back leading to cancellation of writhe between the two helical sections. We include the writhe profiles of the whole molecules for 3F1L and 1KG2 in Figure 5.14 to further highlight that although the dominant geometry of these two appears different, there are many similar subsections picked up by this metric. Indeed, despite 3F1L's dominant globally positive helical geometry, it has many subsections with cancellation of writhe. This in part has led to it clustering with more varied conformations, whereas 1P1X belongs to a much stricter globally helical cluster.

To conclude this subsection, we consider the similarities between the two members of the cluster containing 3F1L that we have highlighted above. The similarity between 3F1L and 1C3D was due to their consistent uniform global helical confor-

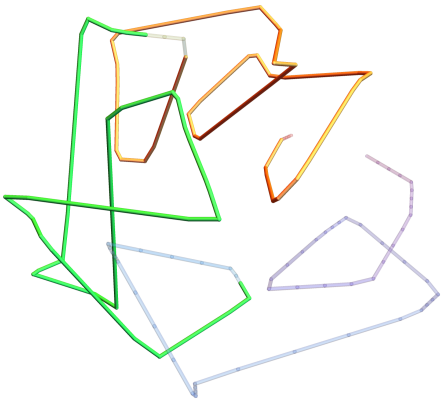
mation. However, the similarity between 3F1L and 1KG2 was due to many smaller helical conformations, in some cases with opposite orientation. To fully understand the characteristics of this cluster, we should investigate then in what way 1C3D and 1KG2 are similar. Their mutual similarity score is 0.68, slightly lower than that of 3F1L and 1KG2. We plot the matched subsections of 1C3D and 1KG2 in Figure 5.15. Of particular note is that the same two subsections of 1KG2 are identified as similar to subsections of 1C3D as in the 3F1L case.

This is even more evident as we plot the matched subsections of the SKMT smoothed backbones of 1C3D and 1KG2 in Figure 5.16. We again see that the first matched subsection is a single clear helical loop, shown in green. The second matched subsection is due to another helical subsection, slightly less uniform in its coiling. Comparing these subsections to those from Figure 5.11, it is clear that the nature of the similarity between 1C3D and 1KG2 is alike to that between 3F1L and 1KG2. In particular, the same subsections of 1KG2 are present in both cases. The similarity between 3F1L and 1C3D is evidently due to these same subsections too, as they are classified as wholly similar. Although the specific arrangement of these subsections may vary between cluster members, there is a clear pattern to the similarities identified which have informed this clustering.

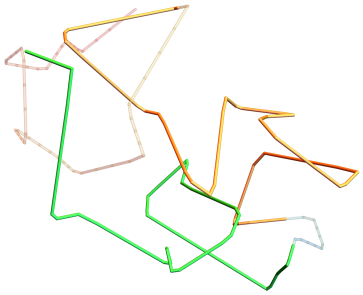
5.3.3 A roadie based cluster

In the previous example, shared helical geometries identified by the writhe metric informed the cluster, with variation due to their uniformity and consistent orientation. As we have seen in Chapter 3, there also exists a super secondary motif consisting of counter helical loops which systematically minimise writhe. The writhe based metric should identify and cluster proteins which contain this conformation. In fact, consider the cluster containing the PDB entry 4NNO, which was seen to have a roadie-like subsection according to the criteria defined in Section 3.4.2. We plot their writhe profiles in Figure 5.17, noting that each of these contains at least one roadie like subsection in their writhe fingerprints.

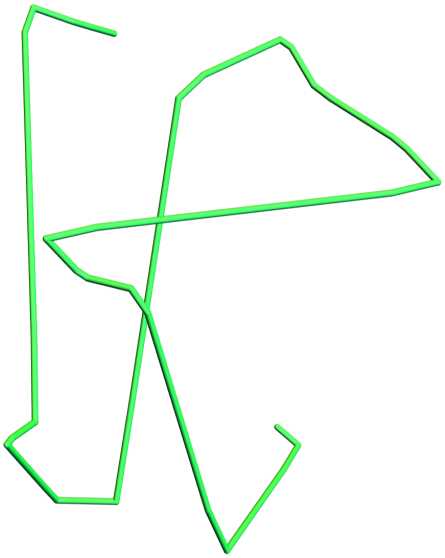
In Figure 5.18 we show the matched subsections of the writhe profiles of 4NNO and another member of its cluster, 7OCB. The metric has identified a roadie like



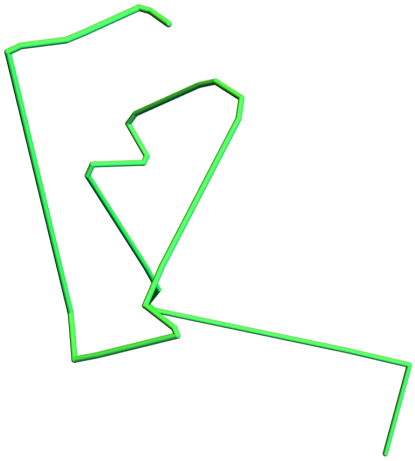
A The SKMT smoothed backbone of 1C3D, with the mutually similar sections to 1KG2 highlighted.



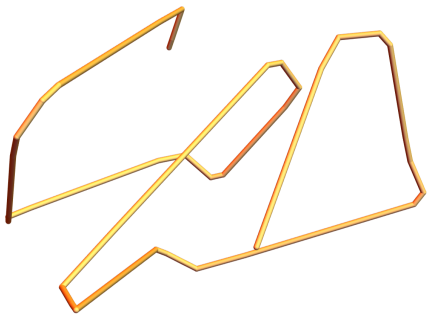
B The SKMT smoothed backbone of 1KG2, with the mutually similar sections to 1C3D highlighted.



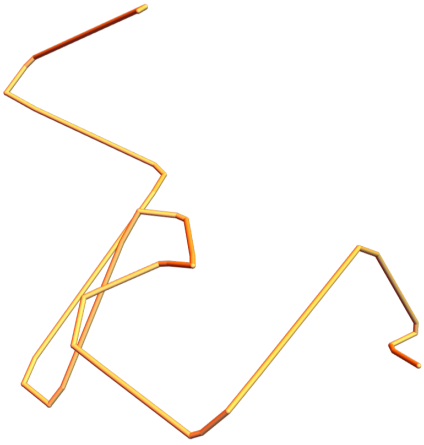
C The first matched subsection of 1C3D.



D The first matched subsection of 1KG2.



E The second matched subsection of 1C3D.



F The second matched subsection of 1KG2.

Figure 5.16: The mutually similar sections of PDB entries 1C3D and 1KG2.

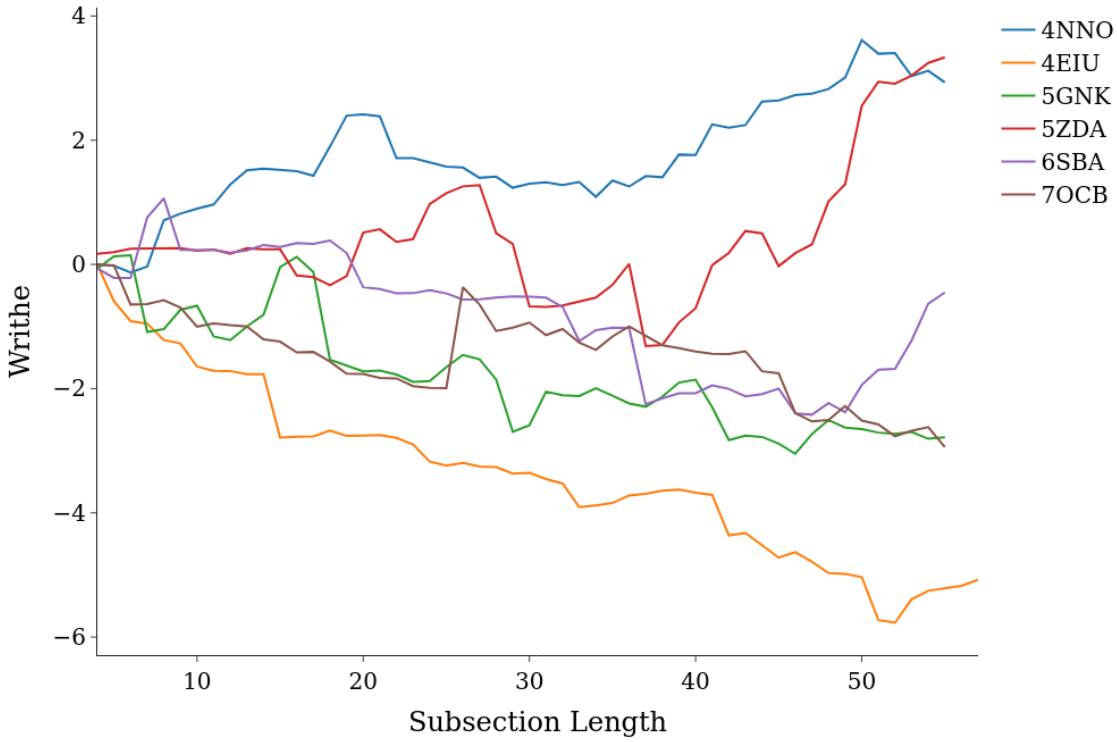


Figure 5.17: The writhe profiles for proteins clustered with PDB entry 4NNO.

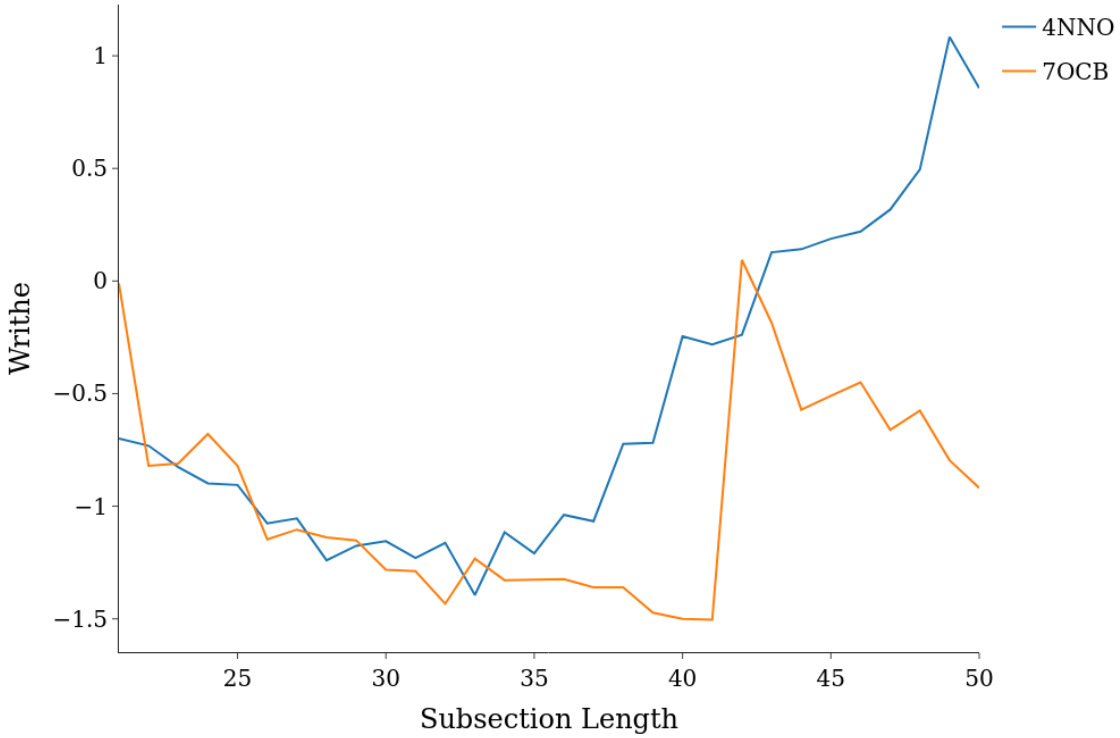
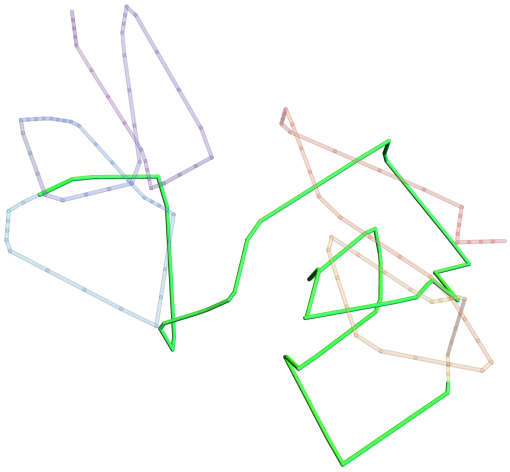
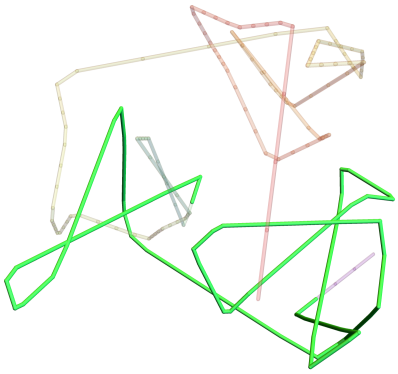


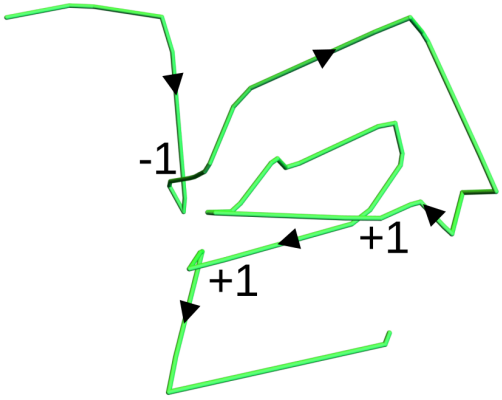
Figure 5.18: The matched subsections of the writhe fingerprints of PDB entries 4NNO and 7OCB.



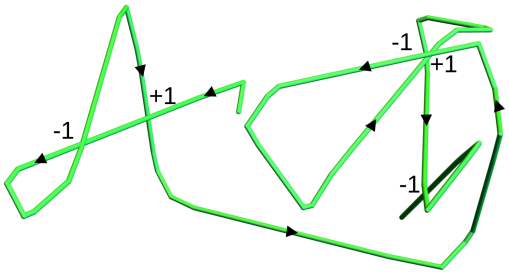
A The SKMT smoothed backbone of 4NNO, with the mutually similar section to 7OCB highlighted.



B The SKMT smoothed backbone of 7OCB, with the mutually similar section to 4NNO highlighted.



C The matched subsection of 4NNO, with significant crossings contributing to the roadie motif annotated.



D The matched subsection of 7OCB, with significant crossings contributing to the roadie motif annotated.

Figure 5.19: The mutually similar roadie subsections of PDB entries 4NNO and 7OCB.

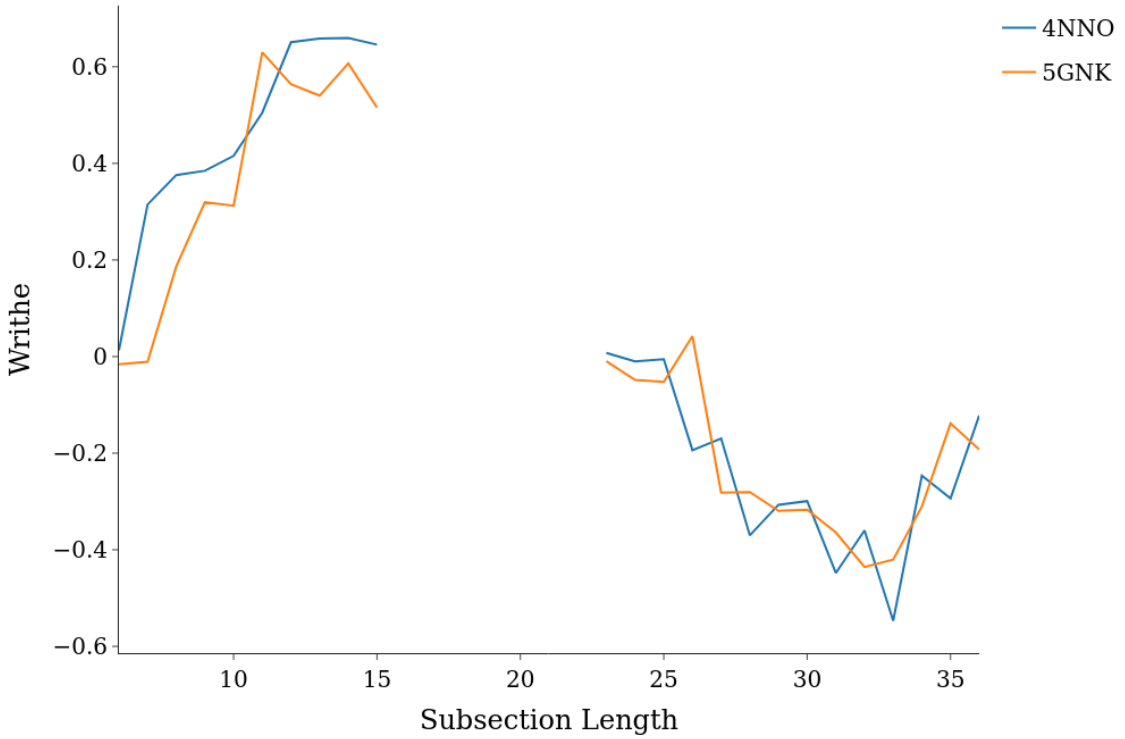
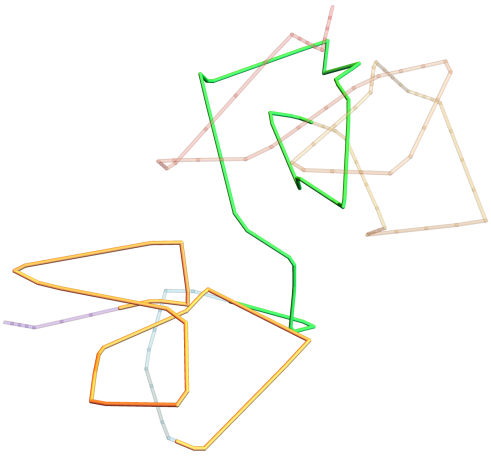


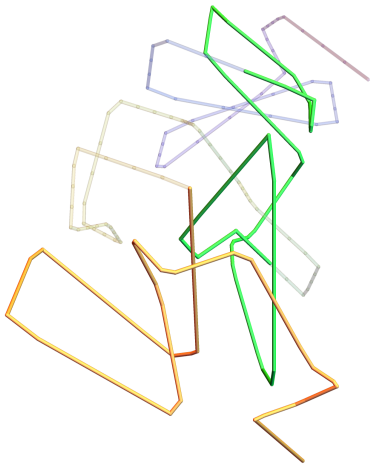
Figure 5.20: The mutually similar subsections of the writhe fingerprints of PDB entries 4NNO and 5GNK.

subsection is common between the two molecules, accounting for 61.8% of their SKMT smoothed backbones. There is a clear deviation after this roadie subsection, with 4NNO continuing to accumulate positive writhe, whereas 7OCB is negatively coiled after the roadie. This can be seen in their SKMT smoothed backbones plotted in Figure 5.19, where the crossings contributing to this roadie motif are annotated. The similarity cut off parameter set at 0.1 allows for similarities such as this where a dominant motif is identified, potentially with deviation either side of it. Where 7OCB has no defined CATH topology, 4NNO consists of two Rossmann Fold domains, with the roadie subsection formed from the ends of each of these two domains. Since the two Rossmann fold domains have this opposite orientation, there is much less writhe build up than seen in 3F1L for example, thus explaining their distinct clustering.

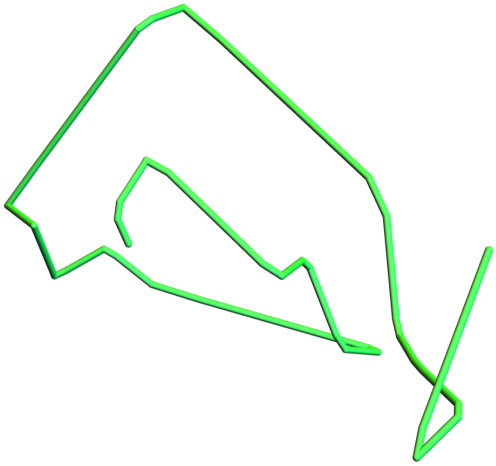
To better understand this cluster, we compare the matched subsections of the writhe fingerprints for 4NNO with another member, 5GNK, in Figure 5.20. In this



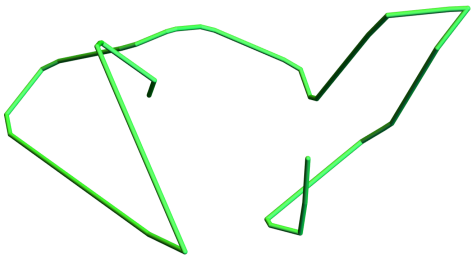
A The SKMT smoothed backbone of 4NNO, with the mutually similar section to 5GNK highlighted.



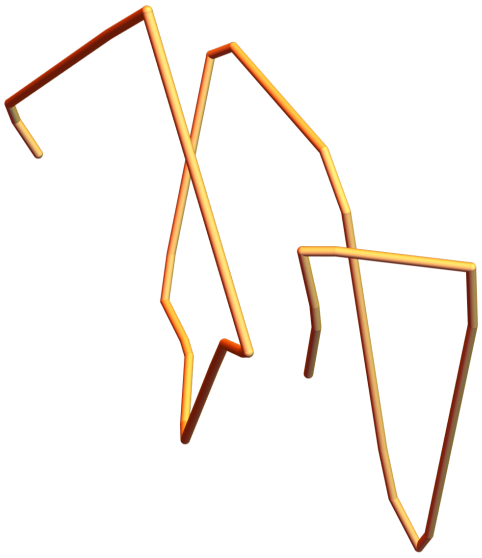
B The SKMT smoothed backbone of 5GNK, with the mutually similar section to 4NNO highlighted.



C The first matched subsection of 4NNO.



D The first matched subsection of 5GNK.



E The second matched subsection of 4NNO.

F The second matched subsection of 5GNK.

Figure 5.21: The mutually similar subsections of PDB entries 4NNO and 5GNK.

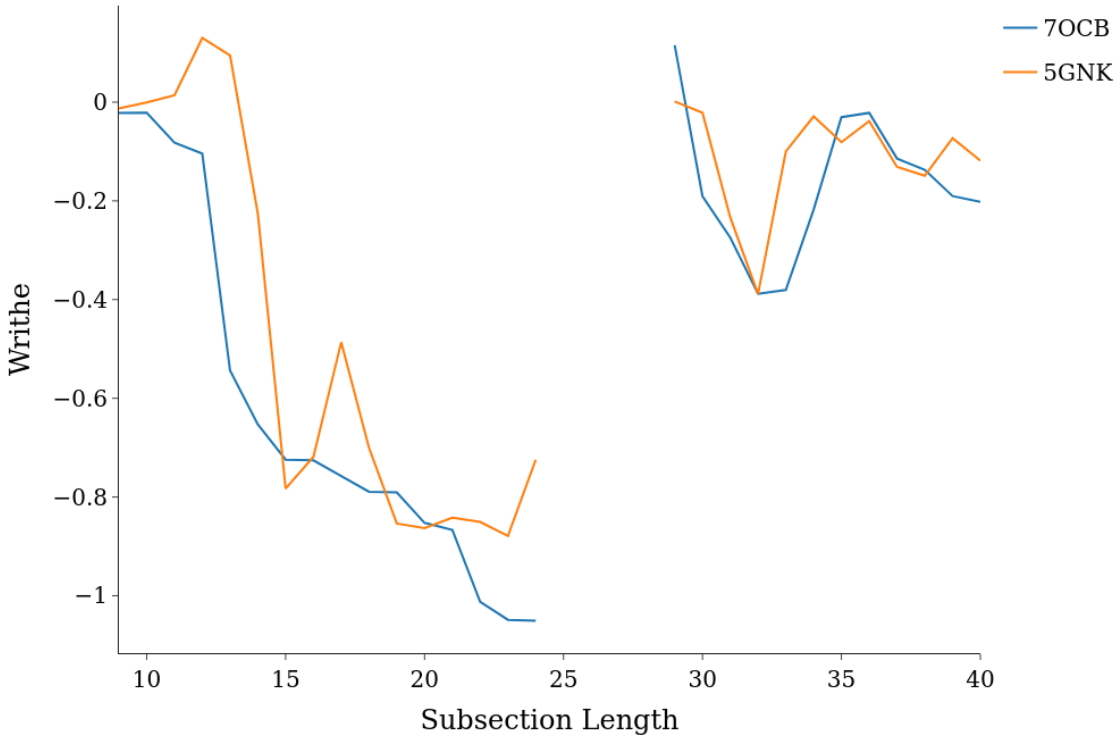


Figure 5.22: The mutually similar subsections of the writhe fingerprints of PDB entries 7OCB and 5GNK.

case, their similarity is due to a much smaller roadie like subsection, combined with a helical subsection. In fact, the matched roadie like subsection from 4NNO comes from the much larger roadie wrap which was shared with 7OCB in the previous example in Figure 5.19. The matched helical subsection from 4NNO is from the first of its two Rossmann Fold domains, with a clearly defined loop. In the case of the more global similarity with 7OCB, this helical subsection's contribution to the similarity was as part of the dominant roadie like geometry. In this case however, it was paired more simply with another helical loop from 5GNK.

To fully characterise this cluster, we will consider the nature of the similarity between 7OCB and 5GNK as they relate to 4NNO. In the case of the 7OCB's similarity with 4NNO, there was a clear large scale roadie wrap conformation similar between the two. For 5GNK, its similarity with 4NNO was again due to this roadie wrap, however a smaller portion of it, combined with shared helical sections elsewhere. In Figure 5.22 we plot the matched subsections of the writhe fingerprints for

7OCB and 5GNK. Of particular note here is that these subsections are again a combination of helical and roadie like motifs. Since a roadie wrap conformation requires helical loops as its building blocks, it is unsurprising that many of the similarities within this cluster are due to shared helical sections. However, its defining feature is the frequent arrangement of these helical subsections which leads to cancellation of writhe. This distinguishes it from say 1P1X's cluster, where the helical subsections are arranged with a consistent orientation leading to the uniform writhe growth and globally helical geometries.

5.3.4 Comparing across clusters

To conclude this section, we will investigate similarities between members of distinct groups to better understand the reasons for their separation. For example, given that 4NNO is composed of two Rossmann Fold domains, and 3F1L is itself a Rossmann Fold, we would imagine that they share a large similarity.

Indeed, we plot the matched subsections of these two proteins in Figure 5.23, with a similarity score of 0.8. The structural similarities are shown in Figure 5.24, with shared helical subsections in green, and a common roadie like conformation in orange. The dominant similarity is due to the helical subsection, as expected from the Rossmann Fold connection between these two. On the other hand, the shared roadie subsection is significantly smaller than the roadie conformations we have seen in 4NNO's cluster.

To completely understand the separation of 3F1L and 4NNO into distinct clusters, we consider the pairwise similarity between some other member of each of their clusters. In Figure 5.25 we plot the matched subsections of the writhe fingerprints of 1C3D, a member of 3F1Ls cluster, and 4EIU, a member of 4NNO's cluster. We see a clear lack of similarity between these two proteins, with a pairwise similarity score of 0.4. The largest similarity between the two is due to small regions of relatively little change in writhe. Considered separately, 1C3D is associated with 3F1L due to a shared large scale positive helical geometry, and, 4EIU with 4NNO due to a shared roadie like subsection. However, we have seen above that 3F1L and 4NNO are highly similar due to the fact they both contain subsections with each of these

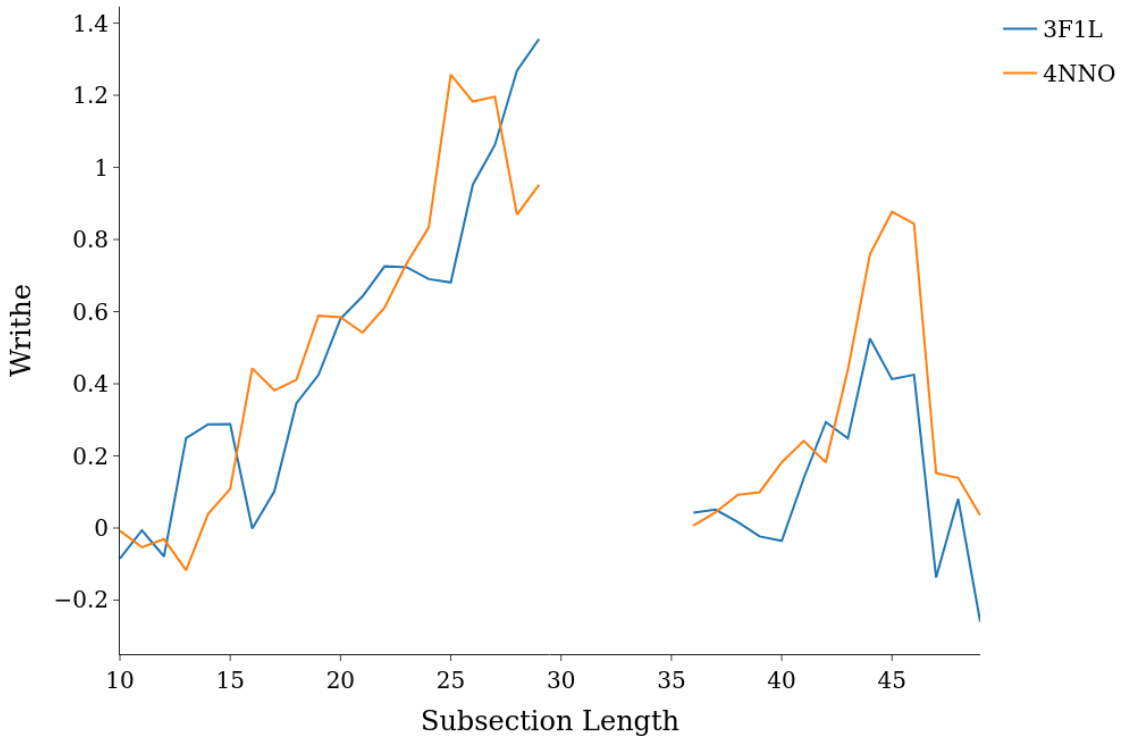
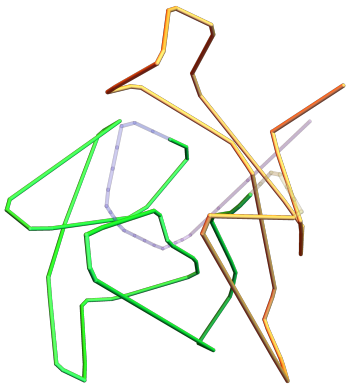
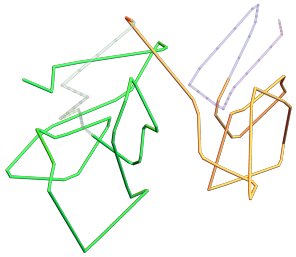


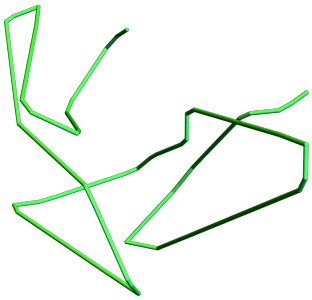
Figure 5.23: The mutually similar subsections of the writhe fingerprints of PDB entries 3F1L and 4NNO.



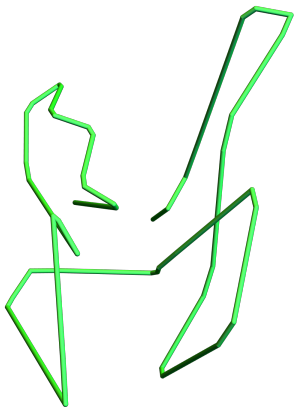
A The SKMT smoothed backbone of 3F1L, with the mutually similar section to 4NNO highlighted.



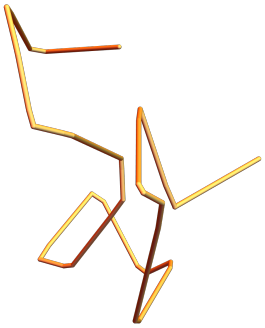
B The SKMT smoothed backbone of 4NNO, with the mutually similar section to 3F1L highlighted.



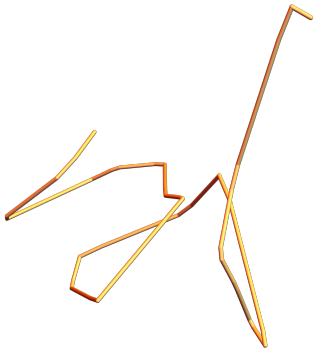
C The first matched subsection of 3F1L.



D The first matched subsection of 4NNO.



E The second matched subsection of 3F1L.



F The second matched subsection of 4NNO.

Figure 5.24: The mutually similar subsections of PDB entries 3F1L and 4NNO.

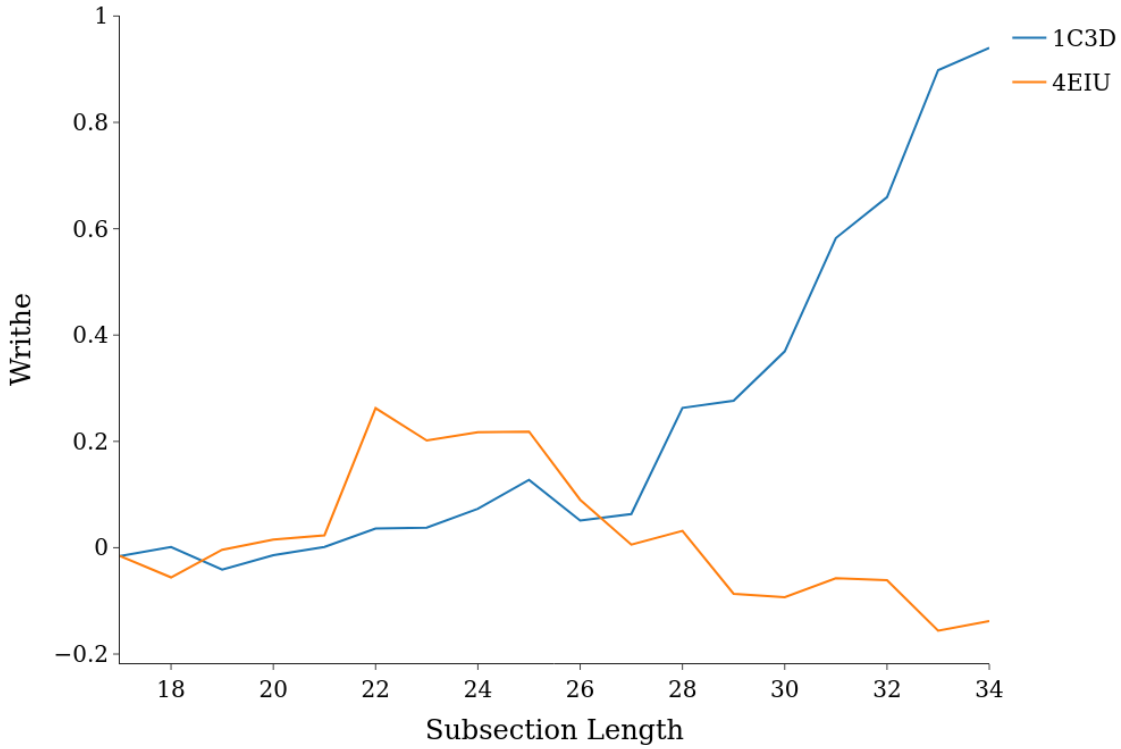


Figure 5.25: The matched subsections of the writhe fingerprints for PDB entries 1C3D and 4EIU

super secondary motifs. Their classification into separate clusters then is due to the relative size and significance of these motifs within their overall structure. In the case of 3F1L, the dominant geometry is helical, with some subsections of writhe cancellation. The most significant geometry for 4NNO however is the large roadie like conformation formed by the connection of its two Rossmann Fold domains.

5.4 Discussion

In this chapter we outlined the need for a similarity metric which is robust to the small scale variations present in low resolution or highly flexible structures. We showed through examples the types of geometric and topological similarities this metric should identify. One aim of the metric that is particular to this work is that it should highlight super secondary motifs that are common between proteins, regardless of the types of secondary structure. This differentiates the metric from other used approaches, identifying similarities that are purely geometric in nature and not affected by potential inaccuracies in the secondary structure.

To meet these aims, we constructed a writhe based similarity metric for the SKMT representation of protein backbones. This metric was defined in two stages. First, we compute the mean absolute difference between the writhe of subsections of the SKMT smoothed backbone curves, relative to the maximum possible difference of their writhes as quantified by the bound from Chapter 3. We then return the largest disjoint subsection(s) where this mean absolute difference is below some user specified tolerance s_0. One potential drawback of this approach is that the output information can be interpreted in multiple ways.

Firstly, we can plot the writhe fingerprints of these subsections as a method of understanding the scale and nature of the entanglements. Alongside this, visual inspection of these subsections of the SKMT smoothed backbones can provide insight into the arrangement of the similar super secondary motifs. As a similarity score, we can also consider the percentage coverage of the SKMT backbones that these subsections represent. Though a simple score between 0 (no similarity) and 1 (identical) may seem desirable, we feel that viewing the different representations of

this metric's output provides a much more complete picture of the similarity. For this reason, we provide some simple examples of applications of this metric for exploratory analysis of a protein database. One limiting factor on the broader use of this metric is an intuitive understanding of both the SKMT curves and the tuning of the parameter s_0. To this end, we provide an interactive iPython notebook where users can follow the examples outlined in this chapter before studying their protein of interest in this manner.

With this writhe metric in place, we performed a complete pairwise sweep across our data set with $s_0 = 0.1$ for clustering. Given the computational complexity of this task, there was a limitation on the time we could devote to exploring these clusters. For this thesis, we reduced the scale by limiting the clustering to proteins within a chosen length range, dictated by the example proteins which motivated the work in this chapter. Within this reduced sample, we found evidence of clusters representing the two super secondary motifs discussed throughout this thesis. In general, we again found helical geometries extremely common across the dataset, with proteins clustered according to the uniformity and orientation of sequential helical loops. The results of our clustering indicate that the helical arrangement of secondary structures is a key step in building tertiary structure.

The depth of information contained within this pairwise sweep is an ideal avenue for further study, with more time one could run a sweep for a lower s_0 to potentially extract tighter clusters or study the persistence of clusters relative to a changing s_0. For instance, we presented an example Rossmann Fold domain whose cluster members could broadly be split into two types. One would imagine that this cluster would indeed split into two clusters for a harsher s_0. Example work in this area (*eg* [50, 61]) associates geometric / topological clusters with CATH classifications or biological features. Although we note the insight these approaches provide, the unique motivation for our work differs somewhat from this aim. That is, our metric was designed to identify structural similarities that may be missed by other approaches, as can be seen in our example clusters containing multiple CATH topologies.

CHAPTER 6

Conclusions and Future Work

The main focus of this thesis has been the development of measures of entanglement for the study of tertiary protein structure, with two main applications. The original motivation for this work was the development of the Carbonara software package; a tool for rapid refinement of protein structure according to BioSAXS data. We show that the topological tools we developed to restrict the search space of this problem have wider applications in classifying protein structures and uncovering surprising structural relationships. Chapter 2 provides the background and context into which the remainder of this thesis fits. In Chapter 2, we introduce the geometrical and topological tools that we will apply to the various scales of protein structure. In Chapter 3, we describe the SKMT algorithm, a novel method for smoothing a backbone curve to provide a minimal representation of the entanglement of secondary structure elements. Computing the writhe of the SKMT smoothed backbone curves, we show that there are clear and consistent length scaling relationships for the entanglement of the backbone curve relative to its secondary structure. Using fundamental features of the writhe, we uncover two distinct supersecondary structural motifs which are common to many proteins. In Chapter 4 we show that there is a minimum expected amount of entanglement for protein backbone curves rela-

tive to their secondary structure. We use this fact to restrict the state space from BioSAXS experiments, a key feature in the development of Carbonara. We then describe some further developments to Carbonara based on the current landscape of tertiary structure prediction. We conclude this chapter by building the foundations of a complementary model to Carbonara. This alternative model takes a gradient descent approach to optimise the backbone curve according to the geometrical and topological measures we have developed in this thesis, along with the BioSAXS data. In Chapter 5, we use the writhe to build a structural similarity metric which is robust to the flexible nature of proteins in solution, and can identify shared topological features. We provide some example use cases for this metric, before applying it to cluster a pairwise similarity sweep of our representative sample of protein backbone curves.

6.1 Chapter 2

We have provided the geometrical and topological background relevant to the study of proteins on the secondary, super secondary, and tertiary structural level. Firstly, we introduce the discrete curvature and torsion which describe the local geometry of backbone curve, and are heavily constrained analogously to the Ramachandran angles. We show that these local geometrical constraints are however not sufficient to describe a globally realistic backbone curve, motivating the development of tools to study tertiary structure. We show through a classical example the structural features that commonly used alignment based comparison methods fail to capture. We introduce the writhe, the topological quantity on which the majority of the work in this thesis is built. We discuss the application of this measure of self entanglement to protein backbone curves, highlighting the key aspects of the literature in this area upon which this thesis provides further insight.

6.2 Chapter 3

With the aim of uncovering the entanglement of the tertiary structure which may be hidden by the local geometry of the secondary structures, we have developed the SKMT algorithm for smoothing the backbone curve. This backbone smoothing method produces a minimal representation, speeding up computation of potentially expensive topological measures, whilst maintaining the recognisable secondary structure arrangements for visual interpretation. With this backbone smoothing in place, we can study the distribution of writhe across the PDB. We find a linear bounding curve captures over 97% of the data. Given the correspondence between helices and linear growth in writhe, this implies that large scale helices provide an entanglement maximising conformation for backbone curves. Moreover, we uncover the presence of helical super secondary structures of various scales, with a writhe gradient similar to that of the global helices. We investigate the strength of this relationship by considering a common feature of experimentally determined structures, namely the potential for missing residues. By studying the distribution of writhe against alternative smoothing methods and definitions of length, we show the SKMT approach provides the clearest representation of the scaling of entanglement.

We conclude this chapter with two brief studies tangential to the central ideas of this thesis. The first is an investigation of a correlation between knottedness of a backbone curve and the writhe of its SKMT smoothed representation. This work helps place the earlier results in this chapter into the context of the wider literature, highlighting the advantage of studying the SKMT backbone representation. In the second, we introduce a method for storing cables which minimises entanglement. By identifying subsections of the writhe profile of SKMT smoothed backbone curves which satisfy the specific writhe criteria of this arrangement, we show that many proteins have subsections adopting this conformation of systematic cancellation of entanglement. The functional advantage of this conformation in other contexts should inspire further study into the reason for their prevalence in proteins, and we include this work to provide the tools for this.

6.3 Chapter 4

We study the distribution of *acn* of SKMT smoothed backbone curves. In particular, we uncover the existence of a minimum expected entanglement of secondary structure elements in the tertiary structure. The inclusion of this length dependent expected entanglement as a penalty in the fitting procedure is a key feature of the development of the constrained backbone algorithm into the Carbonara software. With this penalty in place, we show that Carbonara produces a realistic novel conformation for Human SMARCAL1, based on data collected at Diamond Light Source. We present examples of Carbonara's development relative to the changes in the structural prediction landscape over the course of my PhD. In particular, we show that Carbonara's significant advantage over AlphaFold is its ability to characterise proteins with potentially dynamic sections. This information is only accessible through the realistic interpretation of BioSAXS data present in this model.

The Monte Carlo approach taken in Carbonara is at once its main strength and a limitation. In particular, late in prediction runs we see a significant drop off in the ratio of acceptable changes. To address this, we develop the framework for a complementary Carbonara model which takes a gradient descent approach to optimising protein structures via their BioSAXS data. We presented the foundations of this model with a simple proof-of-concept optimising a low-confidence AlphaFold prediction with disordered regions.

6.4 Chapter 5

We discussed the need for similarity metrics which are robust to small scale conformational changes due to low resolution data or flexibility. To address this need, we construct a writhe based similarity metric which identifies the largest subsections of two curves whose writhe deviates by less than the maximal observed deviation according to the bound from Chapter 3. We showed that this metric can highlight shared helical geometries, as well as identify proteins who are "almost"" knotted but for a change in the threading of their termini. We showed some example applications of this metric, providing an interactive iPython notebook such that researchers can

use these tools.

We applied this similarity metric to construct a pairwise similarity matrix for our complete dataset. As an initial investigation into this complex data, we filtered the sample by length and identified clusters associated with the super secondary motifs which have been a focus throughout this thesis. In particular, we found that proteins can be associated according to the uniformity of their globally helical conformations. In one case, clustering was due to the presence of the roadie wrap conformation, with a protein's two helical Rossmann Fold domains arranged with opposite orientations such that was cancellation of their writhe.

6.5 Potential Future Work

The results of Chapters 3 and 4, namely the bounds on entanglement as measured relative to secondary structure are significant in their improvements to the Carbonara model. To extend these further, it would be useful to determine bounds for both the self and mutual entanglement of proteins with multiple chains. Within this study, for any multimers present in the data set, we considered only their first chain. However, we could remove these examples to be studied separately, thereby strengthening our monomer bound and extending the work. Since Carbonara is capable of making predictions for multimeric proteins in solution, it would be advantageous to understand the range of possible conformations both the individual chains take, as well as the manner in which they entangle with each other.

The exploration of super secondary motifs in Chapter 3, in particular the roadie wrap conformation, is driven primarily by their topological beauty. It is hoped that by providing tools to identify these conformations, this work can be picked by researchers who are better placed to investigate their biological relevance. In particular, experimental studies on the folding rates and pathways of proteins that exhibit these conformations could provide an explanation for their abundance. For example, the folding of proteins on and off the ribosome is investigated in a recent publication [104]. An investigation of the writhe profiles the partially folded intermediates they find on the ribosome could reveal roadie wraps as a key structural

motif in the folding pathway. There has been much work in the structure based protein design using tandem repeat domains which we have shown to be linked to the large scale helical geometries. Given the links between Greek Key motifs and roadie wrap geometries identified in this thesis, there is potential to expand this work into the de novo design of proteins adopting an ideal roadie wrap conformation.

The complementary model to Carbonara presented in Chapter 4 has huge potential to expand the current prediction pipeline. The initial proof of concept shows the significant advantage of this gradient descent approach. The key next step in the development of this model is the inclusion of a penalty such that the curve can be optimised relative to the experimental solution scattering data. This is already in progress, though not in a state to be included in this thesis. Currently, the model will move all linkers simultaneously in the direction that maximises the likelihood according to our defined penalties at each iteration. To move this approach in line with Carbonara, allowing the user to choose which regions are allowed to vary and include any known distance constraints would be obvious next steps. Alongside this, robust testing and tuning of each of the individual penalties, and their relative weighting, are essential.

We were only able to investigate a fraction of the wealth of information contained within the pairwise similarity matrix constructed in Chapter 5. In fact, the clusters determined from this matrix could be used to inform the Carbonara predictions. The better we understand the range of possible conformations that proteins can take relative to their size and secondary structure, the more we can reduce the state-space search for Carbonara. Similar clustering techniques could be implemented as part of the Carbonara pipeline. Given its rapid search approach, we can often have many output predictions with similarly good fits to the scattering data. By clustering these outputs based on the writhe metric, we could highlight the key differences between the predictions to guide further studies such as MD simulations.

The ultimate aspiration of this thesis is that the approach to structural comparison presented, along with the tools and software to perform then, can provide new insight for the community into the rich world of protein structure. With the release of Carbonara as an open source software, we hope to provide an intuitive method

for interpreting BioSAXS which inspires wider adoption of this technique.

Bibliography

[1] C. Tanford and J. Reynolds, *Nature's robots: a history of proteins*. OUP Oxford, 2003. 1.1

[2] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983. 1.1

[3] A. Bale, R. Rambo, and C. Prior, "The skmt algorithm: a method for assessing and comparing underlying protein entanglement," *PLoS Computational Biology*, vol. 19, no. 11, p. e1011248, 2023. 1.2, 3, 5

[4] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu. Rev. Biophys.*, vol. 37, no. 1, pp. 289–316, 2008. 1.3.1, 1.4.2

[5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021. 1.3.1

[6] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021. 1.3.1

[7] B. Zhao, S. Ghadermarzi, and L. Kurgan, "Comparative evaluation of alphafold2 and disorder predictors for prediction of intrinsic disorder, disorder content and fully disordered proteins," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 3248–3258, 2023. 1.3.1

[8] H. Kuboniwa, N. Tjandra, S. Grzesiek, H. Ren, C. B. Klee, and A. Bax, "Solution structure of calcium-free calmodulin," *Nature Structural Biology*, vol. 2, no. 9, pp. 768–776, 1995. 1.2A

[9] R. Chattopadhyaya, W. E. Meador, A. R. Means, and F. A. Quiocho, "Calmodulin structure refined at 1.7 å resolution," *Journal of Molecular Biology*, vol. 228, no. 4, pp. 1177–1192, 1992. 1.2B

[10] D. Chin and A. R. Means, "Calmodulin: a prototypical calcium sensor," *Trends in Cell Biology*, vol. 10, no. 8, pp. 322–328, 2000. 1.3.1

[11] J. Osz, A. G. McEwen, M. Bourguet, F. Przybilla, C. Peluso-Iltis, P. Poussin-Courmontagne, Y. Mély, S. Cianférani, C. M. Jeffries, D. I. Svergun, *et al.*, "Structural basis for dna recognition and allosteric control of the retinoic acid receptors rar–rxr," *Nucleic Acids Research*, vol. 48, no. 17, pp. 9969–9985, 2020. (document), 1.3.2, 1.3

[12] E. Brookes, M. Rocco, P. Vachette, and J. Trewhella, "Alphafold-predicted protein structures and small-angle x-ray scattering: insights from an extended examination of selected data in the small-angle scattering biological data bank," *Journal of Applied Crystallography*, vol. 56, no. 4, pp. 910–926, 2023. 1.3.2

[13] T. Saldaño, N. Escobedo, J. Marchetti, D. J. Zea, J. Mac Donagh, A. J. Velez Rueda, E. Gonik, A. García Melani, J. Novomisky Nechcoff, M. N. Salas, *et al.*, "Impact of protein conformational diversity on alphafold predictions," *Bioinformatics*, vol. 38, no. 10, pp. 2742–2748, 2022. 1.3.2

[14] D. I. Svergun, "Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing," *Biophysical Journal*, vol. 76, no. 6, pp. 2879–2886, 1999. 1.3.2

[15] D. Franke and D. I. Svergun, "Dammif, a program for rapid ab-initio shape determination in small-angle scattering," *Journal of Applied Crystallography*, vol. 42, no. 2, pp. 342–346, 2009. 1.3.2

[16] D. I. Svergun, M. V. Petoukhov, and M. H. Koch, "Determination of domain structure of proteins from x-ray solution scattering," *Biophysical Journal*, vol. 80, no. 6, pp. 2946–2953, 2001. 1.3.2, 4.2.3

[17] R. P. Rambo and J. A. Tainer, "Bridging the solution divide: comprehensive structural analyses of dynamic rna, dna, and protein assemblies by small-angle x-ray scattering," *Current Opinion in Structural Biology*, vol. 20, no. 1, pp. 128–137, 2010. 1.3.2

[18] D. Svergun, C. Barberato, and M. H. Koch, "Crysol–a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates," *Journal of Applied Crystallography*, vol. 28, no. 6, pp. 768–773, 1995. 1.3.2, 4.1.2, 4.1.2

[19] D. Schneidman-Duhovny, M. Hammel, and A. Sali, "Foxs: a web server for rapid computation and fitting of saxs profiles," *Nucleic Acids Research*, vol. 38, no. suppl_2, pp. W540–W544, 2010. 1.3.2, 4.1.2, 4.1.2, 4.2.3

[20] F. Poitevin, H. Orland, S. Doniach, P. Koehl, and M. Delarue, "Aquasaxs: a web server for computation and fitting of saxs profiles with non-uniformally hydrated atomic models," *Nucleic Acids Research*, vol. 39, no. suppl_2, pp. W184–W189, 2011. 1.3.2

[21] P.-c. Chen and J. S. Hub, "Interpretation of solution x-ray scattering by explicit-solvent molecular dynamics," *Biophysical Journal*, vol. 108, no. 10, pp. 2573–2584, 2015. 1.3.2

[22] H. J. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, vol. 91, no. 1-3, pp. 43–56, 1995. 1.3.2

[23] C. Prior, O. R. Davies, D. Bruce, and E. Pohl, "Obtaining tertiary protein structures by the ab initio interpretation of small angle x-ray scattering data," *Journal of Chemical Theory and Computation*, vol. 16, no. 3, pp. 1985–2001, 2020. 1.3.2, 1.4, 2.1.4, 2.2.2, 2.5, 4.1.1, 4.1, 4.1.2, 4.1.2, 4.2.1, 4.2.3, 4.2.4

[24] R. P. Rambo and J. A. Tainer, "Accurate assessment of mass, models and resolution by small-angle scattering," *Nature*, vol. 496, no. 7446, pp. 477–481, 2013. 1.3.3

[25] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. M. Pang, L. Woodridge, C. Rauer, N. Sen, M. Abbasian, S. Le Cornu, S. D. Lam, K. Berka, I. Varekova, R. Svobodova, J. Lees, and C. A. Orengo, "CATH: increased structural coverage of functional space," *Nucleic Acids Research*, vol. 49, pp. D266–D273, 11 2020. 1.4.1

[26] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "Scop: a structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536–540, 1995. 1.4.1

[27] M. Vijayabaskar and S. Vishveshwara, "Insights into the fold organization of tim barrel from interaction energy based structure networks," *PLOS Computational Biology*, vol. 8, no. 5, p. e1002505, 2012. 1.4.2

[28] M. Figueroa, M. Sleutel, M. Vandevenne, G. Parvizi, S. Attout, O. Jacquin, J. Vandenameele, A. W. Fischer, C. Damblon, E. Goormaghtigh, *et al.*, "The unexpected structure of the designed protein octarellin v. 1 forms a challenge for protein structure prediction tools," *Journal of Structural Biology*, vol. 195, no. 1, pp. 19–30, 2016. 1.4.2, 5.1

[29] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, pp. 585–590, 1977. 1.4.2

[30] M. Karplus and D. L. Weaver, "Diffusion–collision model for protein folding," *Biopolymers: Original Research on Biomolecules*, vol. 18, no. 6, pp. 1421–1437, 1979. 1.4.2

[31] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, and Y. Zhou, "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?," *Briefings in Bioinformatics*, vol. 19, no. 3, pp. 482–494, 2018. 1.5

[32] C. Ramakrishnan and G. Ramachandran, "Stereochemical criteria for polypeptide and protein chain conformations: Ii. allowed conformations for a pair of peptide units," *Biophysical Journal*, vol. 5, no. 6, pp. 909–933, 1965. 2.1.1

[33] T. F. Banchoff and S. Lovett, *Differential geometry of curves and surfaces*. Chapman and Hall/CRC, 2022. 2.1.2

[34] A. Hausrath and A. Goriely, "Continuous representations of proteins: Construction of coordinate models from curvature profiles," *Journal of Structural Biology*, vol. 158, no. 3, pp. 267–281, 2007. 2.1.4

[35] A. Krokhotin, A. J. Niemi, and X. Peng, "Soliton concepts and protein structure," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 85, no. 3, p. 031906, 2012. 2.1.4

[36] M. Sadowski and G. Roberts, "Protein structure comparison methods," *Encycl. Biophys.*, pp. 2055–2060, 2013. 2.2.1

[37] L. Holm, "Dali server: structural unification of protein families," *Nucleic Acids Research*, vol. 50, no. W1, pp. W210–W215, 2022. 2.2.1

[38] Y. Zhang and J. Skolnick, "Tm-align: a protein structure alignment algorithm based on the tm-score," *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, 2005. 2.2.1

[39] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 4, pp. 702–710, 2004. 2.2.1

[40] Z. Li, L. Jaroszewski, M. Iyer, M. Sedova, and A. Godzik, "Fatcat 2.0: towards a better understanding of the structural diversity of proteins," *Nucleic Acids Research*, vol. 48, no. W1, pp. W60–W64, 2020. 2.2.1

[41] M. Mansfield, "Are there knots in proteins?," *Nature Structural Biology*, vol. 1, pp. 213–4, 05 1994. 2.2.2, 2.2.3

[42] P. Dabrowski-Tumanski and J. I. Sulkowska, "Topological knots and links in proteins," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3415–3420, 2017. 2.2.2

[43] M. L. Mansfield, "Fit to be tied," *Nature Structural Biology*, vol. 4, no. 3, pp. 166–167, 1997. 2.2.3

[44] W. R. Taylor, "A deeply knotted protein structure and how it might fold," *Nature*, vol. 406, no. 6798, pp. 916–919, 2000. 2.2.3, 3.1.1, 3.1.2, 3.5

[45] P. Dabrowski-Tumanski, P. Rubach, D. Goundaroulis, J. Dorier, P. Sułkowski, K. C. Millett, E. J. Rawdon, A. Stasiak, and J. I. Sulkowska, "Knotprot 2.0: a database of proteins with knots and other entangled structures," *Nucleic Acids Research*, vol. 47, no. D1, pp. D367–D375, 2019. (document), 2.8, 2.2.3, 2.3

[46] N. P. King, E. O. Yeates, and T. O. Yeates, "Identification of rare slipknots in proteins and their implications for stability and folding," *Journal of Molecular Biology*, vol. 373, no. 1, pp. 153–166, 2007. 2.2.3

[47] M. Jamroz, W. Niemyska, E. J. Rawdon, A. Stasiak, K. C. Millett, P. Sułkowski, and J. I. Sulkowska, "Knotprot: a database of proteins with knots and slipknots," *Nucleic Acids Research*, vol. 43, no. D1, pp. D306–D314, 2015. 2.2.3, 5.2.4

[48] V. Turaev, "Knotoids," *Osaka Journal of Mathematics*, vol. 49, no. 1, pp. 195 – 223, 2012. 2.2.3

[49] D. Goundaroulis, J. Dorier, F. Benedetti, and A. Stasiak, "Studies of global and local entanglements of individual protein chains using the concept of knotoids," *Scientific Reports*, vol. 7, no. 1, p. 6309, 2017. 2.2.3

[50] A. Barbensi, N. Yerolemou, O. Vipond, B. I. Mahler, P. Dabrowski-Tumanski, and D. Goundaroulis, "A topological selection of folding pathways from native states of knotted proteins," *Symmetry*, vol. 13, no. 9, p. 1670, 2021. (document), 2.2.3, 3.4.1, 3.16, 5.4

[51] G. Calugareanu, "L'intégrale de gauss et l'analyse des nœuds tridimensionnels," *Rev. Math. Pures et Appl.*, vol. 4, no. 5, 1959. 2.3

[52] R. L. Ricca and B. Nipoti, "Gauss'linking number revisited," *Journal of Knot Theory and Its Ramifications*, vol. 20, no. 10, pp. 1325–1343, 2011. 2.3

[53] K. Klenin and J. Langowski, "Computation of writhe in modeling of supercoiled dna," *Biopolymers: Original Research on Biomolecules*, vol. 54, no. 5, pp. 307–317, 2000. 2.3, 2.3

[54] Y. A. Fosado, D. Michieletto, C. A. Brackley, and D. Marenduzzo, "Nonequilibrium dynamics and action at a distance in transcriptionally driven dna supercoiling," *Proceedings of the National Academy of Sciences*, vol. 118, no. 10, p. e1905215118, 2021. 2.3.1

[55] J. Smrek, J. Garamella, R. Robertson-Anderson, and D. Michieletto, "Topological tuning of dna mobility in entangled solutions of supercoiled plasmids," *Science Advances*, vol. 7, no. 20, p. eabf9260, 2021. 2.3.1

[56] D. Sumners, "The role of knot theory in dna research," in *Geometry and Topology*, pp. 297–318, CRC Press, 2020. 2.3.1

[57] M. Levitt, "Protein folding by restrained energy minimization and molecular dynamics," *Journal of Molecular Biology*, vol. 170, no. 3, pp. 723–764, 1983. 2.3.1

[58] G. A. Arteca and O. Tapia, "Characterization of fold diversity among proteins with the same number of amino acid residues," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 4, pp. 642–649, 1999. 2.3.1

[59] G. A. Arteca, "Scaling regimes of molecular size and self-entanglements in very compact proteins," *Physical Review E*, vol. 51, no. 3, p. 2600, 1995. 2.3.1, 2.3.3

[60] P. Røgen and H. Bohr, "A new family of global protein shape descriptors," *Mathematical Biosciences*, vol. 182, no. 2, pp. 167–181, 2003. 2.3.1

[61] P. Røgen and B. Fain, "Automatic classification of protein structure by using gauss integrals," *Proceedings of the National Academy of Sciences*, vol. 100, no. 1, pp. 119–124, 2003. 2.3.1, 2.4.2, 5.4

[62] P. L. Chang, A. W. Rinne, and T. G. Dewey, "Structure alignment based on coding of local geometric measures," *BMC bioinformatics*, vol. 7, pp. 1–10, 2006. 2.3.1

[63] D. Zhi, M. Shatsky, and S. E. Brenner, "Alignment-free local structural search by writhe decomposition," *Bioinformatics*, vol. 26, no. 9, pp. 1176–1184, 2010. 2.3.1, 3.1.1

[64] C. Grønbæk, T. Hamelryck, and P. Røgen, "Gisa: using gauss integrals to identify rare conformations in protein structures," *PeerJ*, vol. 8, p. e9159, 2020. 2.3.1

[65] J. L. Sleiman, F. Conforto, Y. A. G. Fosado, and D. Michieletto, "Geometric learning of knot topology," *Soft Matter*, vol. 20, no. 1, pp. 71–78, 2024. 2.3.2

[66] J. H. White and W. R. Bauer, "Calculation of the twist and the writhe for representative models of dna," *Journal of Molecular Biology*, vol. 189, no. 2, pp. 329–341, 1986. 2.3.2

[67] E. Orlandini, M. Tesi, S. Whittington, D. Sumners, and E. J. Van Rensburg, "The writhe of a self-avoiding walk," *Journal of Physics A: Mathematical and General*, vol. 27, no. 10, p. L333, 1994. 2.3.3

[68] G. A. Arteca, "Scaling behavior of some molecular shape descriptors of polymer chains and protein backbones," *Physical Review E*, vol. 49, no. 3, p. 2417, 1994. 2.3.3, 3.2

[69] G. A. Arteca, "Self-similarity in entanglement complexity along the backbones of compact proteins," *Physical Review E*, vol. 56, no. 4, p. 4516, 1997. 2.3.3, 4.3.7

[70] J. Cantarella, D. DeTurck, and H. Gluck, "Upper bounds for the writhing of knots and the helicity of vector fields," *AMS IP Studies in Advanced Mathematics*, vol. 24, pp. 1–22, 2001. 2.3.3, 1, 3.2, 3.5

[71] M. A. Berger and G. B. Field, "The topological properties of magnetic helicity," *Journal of Fluid Mechanics*, vol. 147, pp. 133–148, 1984. 2.3.3

[72] J. R. Banavar, A. Maritan, C. Micheletti, and A. Trovato, "Geometry and physics of proteins," *Proteins: Structure, Function, and Bioinformatics*, vol. 47, no. 3, pp. 315–322, 2002. 2.3.3, 1

[73] Y. Diao, A. Dobay, R. B. Kusner, K. Millett, and A. Stasiak, "The average crossing number of equilateral random polygons," *Journal of Physics A: Mathematical and General*, vol. 36, no. 46, p. 11561, 2003. (document), 2.3.3, 4.2, 4.2.1

[74] A. Dobay, J. Dubochet, A. Stasiak, and Y. Diao, "Scaling of the average crossing number in equilateral random walks, knots and proteins," in *Physical And Numerical Models In Knot Theory: Including Applications to the Life Sciences*, pp. 219–231, World Scientific, 2005. 2.3.3, 4.2.1

[75] E. Panagiotou, K. C. Millett, and S. Lambropoulou, "The linking number and the writhe of uniform random walks and polygons in confined spaces," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, no. 4, p. 045208, 2010. 2.3.3

[76] P. Smith and E. Panagiotou, "The second vassiliev measure of uniform random walks and polygons in confined space," *Journal of Physics A: Mathematical and Theoretical*, vol. 55, no. 9, p. 095601, 2022. 2.3.3

[77] D. Hinds and M. Levitt, "A lattice model for protein structure prediction at low resolution.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 7, pp. 2536–2540, 1992. 2.4.2

[78] V. N. Maiorov and G. M. Crippen, "Size-independent comparison of protein three-dimensional structures," *Proteins: Structure, Function, and Bioinformatics*, vol. 22, no. 3, pp. 273–283, 1995. 2.4.2

[79] K. Koniaris and M. Muthukumar, "Knottedness in ring polymers," *Physical Review Letters*, vol. 66, no. 17, p. 2211, 1991. 2.4.2, 3.1.1, 3.1.2, 3.5

[80] K. Lindorff-Larsen, P. Røgen, E. Paci, M. Vendruscolo, and C. M. Dobson, "Protein folding and the organization of the protein topology universe," *Trends in biochemical sciences*, vol. 30, no. 1, pp. 13–19, 2005. 2.4.2, 3.1.1

[81] P. Røgen, "Evaluating protein structure descriptors and tuning gauss integral based descriptors," *Journal of Physics: Condensed Matter*, vol. 17, no. 18, p. S1523, 2005. 2.4.2

[82] L. J. McGuffin, K. Bryson, and D. T. Jones, "The psipred protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000. 3.1.2

[83] R. J. Segura and F. R. Feito, "An algorithm for determining intersection segment-polygon in 3d," *Computers & Graphics*, vol. 22, no. 5, pp. 587–592, 1998. 3.1.2

[84] T. O. Yeates, T. S. Norcross, and N. P. King, "Knotted and topologically complex proteins as models for studying folding and stability," *Current Opinion in Chemical Biology*, vol. 11, no. 6, pp. 595–603, 2007. 3.2

[85] A. V. Efimov, "Super-secondary structures and modeling of protein folds," *Protein Supersecondary Structures*, pp. 177–189, 2013. 3.2.2

[86] I. Hanukoglu, "Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites," *Biochemistry and Molecular Biology Education*, vol. 43, no. 3, pp. 206–209, 2015. 3.2.2

[87] R. Wierenga, "The tim-barrel fold: a versatile framework for efficient enzymes," *FEBS Letters*, vol. 492, no. 3, pp. 193–198, 2001. 3.2.2

[88] L. Paladin, M. Bevilacqua, S. Errigo, D. Piovesan, I. Mičetić, M. Necci, A. M. Monzon, M. L. Fabre, J. L. Lopez, J. F. Nilsson, *et al.*, "Repeatsdb in 2021: improved data and extended classification for protein tandem repeat structures," *Nucleic Acids Research*, vol. 49, no. D1, pp. D452–D457, 2021. 3.2.2

[89] R. C. Lua and A. Y. Grosberg, "Statistics of knots, geometry of conformations, and evolution of proteins," *PLOS Computational Biology*, vol. 2, no. 5, p. e45, 2006. 3.4.1

[90] Z. Haratipour, H. Aldabagh, Y. Li, and L. H. Greene, "Network connectivity, centrality and fragmentation in the greek-key protein topology," *The Protein Journal*, vol. 38, pp. 497–505, 2019. 3.4.2

[91] E. G. Hutchinson and J. M. Thornton, "The greek key motif: extraction, classification and analysis," *Protein Engineering, Design and Selection*, vol. 6, no. 3, pp. 233–245, 1993. 3.4.2

[92] C. D. Putnam, M. Hammel, G. L. Hura, and J. A. Tainer, "X-ray solution scattering (saxs) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution," *Quarterly Reviews of Biophysics*, vol. 40, no. 3, pp. 191–285, 2007. 4.1.2

[93] P. Debye, "Zerstreuung von röntgenstrahlen," *Annalen der Physik*, vol. 351, no. 6, pp. 809–823, 1915. 4.1.2, 4.1.2

[94] C. L. Farrow and S. J. Billinge, "Relationship between the atomic pair distribution function and small-angle scattering: implications for modeling of nanoparticles," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 65, no. 3, pp. 232–239, 2009. 4.1.2

[95] R. Schwartz, S. Istrail, and J. King, "Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues," *Protein Science*, vol. 10, no. 5, pp. 1023–1031, 2001. 4.1.2

[96] D. Svergun, S. Richard, M. Koch, Z. Sayers, S. Kuprin, and G. Zaccai, "Protein hydration in solution: experimental observation by x-ray and neutron scattering," *Proceedings of the National Academy of Sciences*, vol. 95, no. 5, pp. 2267–2272, 1998. 4.1.2

[97] M. A. Coleman, J. A. Eisen, and H. W. Mohrenweiser, "Cloning and characterization of harp/smarcal1: a prokaryotic hepa-related snf2 helicase protein from human and mouse," *Genomics*, vol. 65, no. 3, pp. 274–282, 2000. 4.2.3

[98] G. L. Hura, A. L. Menon, M. Hammel, R. P. Rambo, F. L. Poole Ii, S. E. Tsutakawa, F. E. Jenney Jr, S. Classen, K. A. Frankel, R. C. Hopkins, *et al.*, "Robust, high-throughput solution structural analyses by small angle x-ray scattering (saxs)," *Nature Methods*, vol. 6, no. 8, pp. 606–612, 2009. 4.2.3

[99] S. Chakraborty, R. Venkatramani, B. J. Rao, B. Asgeirsson, and A. M. Dandekar, "Protein structure quality assessment based on the distance profiles of consecutive backbone cα atoms," *F1000Research*, vol. 2, 2013. 4.3.3

[100] P. Røgen, "Quantifying steric hindrance and topological obstruction to protein structure superposition," *Algorithms for Molecular Biology*, vol. 16, pp. 1–19, 2021. 4.3.3

[101] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall III, J. Snoeyink, J. S. Richardson, *et al.*, "Molprobity: all-atom contacts and structure validation for proteins and nucleic acids," *Nucleic Acids Research*, vol. 35, no. suppl_2, pp. W375–W383, 2007. 4.3.8

[102] D. M. Chudakov, M. V. Matz, S. Lukyanov, and K. A. Lukyanov, "Fluorescent proteins and their applications in imaging living cells and tissues," *Physiological Reviews*, vol. 90, no. 3, pp. 1103–1163, 2010. 5.2.2

[103] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172, Springer, 2013. 5.3.2

[104] J. O. Streit, I. V. Bukvin, S. H. Chan, S. Bashir, L. F. Woodburn, T. Włodarski, A. M. Figueiredo, G. Jurkeviciute, H. K. Sidhu, C. R. Hornby, *et al.*, "The ribosome lowers the entropic penalty of protein folding," *Nature*, pp. 1–8, 2024. 6.5